Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Crisis-Aware Variational Autoencoders for Stress Testing Volatility Surfaces

Bienvenue Feugang Nteumagne

feuganteu@gmail.com

University of Johannesburg

Hermann Azemtsa Donfack

University of Johannesburg

Celestin Wafo Soh

Jackson State University

Research Article

Keywords: Variational Autoencoders, Stress Testing, Volatility Surfaces, Exotic Derivatives, Risk Management, Crisis Modelling, Heavy-Tailed Distributions, Arbitrage-Free Constraints

Posted Date: July 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10308753/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Crisis-Aware Variational Autoencoders for Stress Testing Volatility Surfaces

B Feugang Nteumagné[1], H Azemtsa Donfack[1], and C Wafo Soh[1,2,*]
[1]Department of Finance and Investment Management, University of Johannesburg,
Johannesburg 2038, South Africa
[2]Department of Mathematics and Statistical Sciences, College of Science, Engineering and Technology, Jackson State University, Jackson, MS 39217, USA

July 10, 2026

Abstract

We introduce a Crisis-Aware Variational Autoencoder (Crisis VAE) for stress testing volatility surfaces in exotic derivatives pricing. Unlike methods that resample historical data or apply uniform shocks, Crisis VAE combines explicit crisis conditioning through a five-dimensional crisis indicator, a heavy-tailed Student-t latent prior, and soft arbitrage-free constraints to generate novel, internally consistent stressed surfaces with learned smile and term-structure dynamics rather than parallel shifts. Trained on 20 000 Heston-calibrated surfaces spanning normal, stressed, and crisis regimes (with robustness confirmed on Bates jump-diffusion surfaces), the model maps crisis severity directly to plausible, arbitrage-free deformations.

Pricing a portfolio of down-and-in barrier options, target redemption forwards, and variance swaps from the generated surfaces, Crisis VAE produces the tightest, most controllable stress estimates among four established benchmarks (barrier P&L standard deviation \$0.37 versus \$1.50 for historical simulation and \$1.43 for GARCH; 95% minimum holder gain \$6.67 versus \$4.92 and \$4.94). We further show that the parametric delta–gamma approximation fails through *sign reversal*, predicting that a down-and-in put can lose value under volatility stress, contradicting the monotonicity confirmed by full repricing. The framework yields distributional risk measures (VaR, expected shortfall) under explicit crisis-severity control, and we benchmark the generated surfaces structurally against real S&P 500 option data with an out-of-sample surface reconstruction. The method ranking itself is established on synthetic data, where ground truth is controlled; full real-market validation and independent model review remain prerequisites for regulatory deployment.

Keywords: Variational Autoencoders, Stress Testing, Volatility Surfaces, Exotic Derivatives, Risk Management, Crisis Modelling, Heavy-Tailed Distributions, Arbitrage-Free Constraints

1 Introduction

Stress testing of volatility surfaces constitutes a fundamental challenge in derivatives risk management. Financial institutions must assess portfolio losses under severe but plausible market conditions, yet existing methodologies face critical limitations when applied to exotic derivatives whose values depend on the entire volatility surface rather than individual points. The complexity is amplified by the need to maintain internal consistency, smile dynamics, term structure relationships, and arbitrage-free conditions, across stressed surfaces.

Building on the foundational VAE methodology for volatility surface modeling established in our recent work [15], this paper introduces a Crisis-Aware VAE framework specifically designed for stress testing applications. The Crisis VAE extends standard generative modeling by incorporating explicit crisis conditioning, heavy-tailed latent priors, and arbitrage-aware training penalties that produce realistic extreme market conditions while preserving economic consistency.

Traditional stress testing approaches each suffer from fundamental limitations. Historical simulation [4] is constrained to previously observed market states and cannot generate novel crisis scenarios. Parametric approaches [2] apply uniform shocks (parallel shifts, twist, butterfly) that destroy the internal structure of volatility surfaces. GARCH-based methods [2] capture volatility clustering but model only the level of volatility, not its surface shape. Expert-designed scenarios [29] provide single point estimates without distributional risk measures.

The challenge becomes particularly acute for exotic derivatives. Target Redemption Forwards (TARFs), Variance Swaps, and Barrier Options exhibit non-linear payoff structures with sensitivity to different regions of the volatility surface. A parallel vol shift that adequately stresses a vanilla option portfolio may completely misrepresent the risk to a barrier option whose value depends critically on skew dynamics near the barrier level. Variance swaps, through their replication via strips of options across all strikes, are sensitive to the entire surface shape, not just the ATM level.

Our Crisis VAE addresses these limitations through three key innovations. First, a five-dimensional crisis indicator vector $\mathbf{c}$ conditions both the prior distribution and the decoder, enabling explicit control over crisis severity and type. Second, a Student-t prior with $\nu = 4$ degrees of freedom in the latent space encourages generation of fat-tailed scenarios that match the empirical behavior of financial markets during stress periods. Third, soft differentiable penalties for calendar spread and butterfly arbitrage violations ensure generated surfaces remain economically meaningful.

We demonstrate the framework's effectiveness through systematic comparison against four established stress testing methods, pricing three exotic instruments across 500 stress scenarios per method. The results establish Crisis VAE as the most comprehensive approach for volatility surface stress testing, producing the tightest and most conservative risk estimates while maintaining full surface consistency and distributional output.

Specifically, this paper addresses four research questions:

RQ1. Can a crisis-conditioned VAE generate realistic, arbitrage-consistent stressed volatility surfaces with controllable severity?

RQ2. Does full-surface stress testing produce materially different risk estimates for exotic derivatives compared to parallel-shift methods?

RQ3. Is the method ranking robust to the choice of training distribution (Heston vs. Bates jump-diffusion)?

RQ4. Does the latent space capture economically meaningful surface shape variation beyond the crisis severity level?

2 Literature Review and Motivation

2.1 Regulatory Context for Stress Testing

Stress testing in the regulatory sense refers to the evaluation of a financial institution's resilience under hypothetical adverse scenarios, as mandated by the Basel Committee on Banking Supervision [5]. Under the Fundamental Review of the Trading Book (FRTB), banks using Internal Models Approaches (IMA) must compute Expected Shortfall at the 97.5% confidence level over

a base horizon of 10 days, incorporating stressed calibration windows and non-modellable risk factors [6]. Crucially, the FRTB requires that stressed scenarios be *forward-looking* and *plausible*, not merely historical replays, motivating generative approaches that can produce novel crisis conditions while maintaining economic consistency. The Crisis VAE is designed with these requirements in mind, generating distributional stress scenarios with explicit severity control. Formal alignment with FRTB's specific stressed-window calibration requirements would require validation on actual historical market data and regulatory review, which remain priorities for future work.

2.2 Limitations of Traditional Stress Testing

Traditional stress testing methodologies suffer from interconnected limitations that constrain their effectiveness for exotic derivatives portfolios. Volatility-Weighted Historical Simulation (VWHS), despite its success in capturing volatility clustering [4], remains fundamentally bounded by historical data. It cannot generate crisis scenarios that exceed observed market conditions, a critical limitation given that each financial crisis exhibits novel characteristics.

GARCH-based approaches [2] model conditional variance dynamics through the recursion $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$, capturing volatility persistence and mean reversion. However, these models impose a single variance process on the entire surface, reducing the rich structure of strike-maturity dynamics to a scalar. Uncapped GARCH(1,1) forecasts can produce unrealistic terminal volatilities due to the multiplicative nature of the variance process over 252 daily steps. In our implementation we apply a variance cap at the 99.5th percentile of training volatility levels (approximately 41.2%), which constrains forecasts to the support of the training distribution while preserving GARCH dynamics below the cap.

Parametric (delta-gamma) approaches approximate P&L through sensitivity expansions: $\text{P\&L} \approx \text{Vega} \cdot \Delta\sigma + \frac{1}{2}\text{Volga} \cdot \Delta\sigma^2$. While computationally efficient, this approximation breaks down catastrophically for path-dependent exotics. Our experiments show that the delta-gamma approximation produces a sign reversal for the barrier option: it predicts that the Down-and-In Put can lose value under some stress scenarios, despite the theoretical guarantee that higher implied volatility strictly increases a knock-in put's value.

Expert-designed scenarios [29] provide economically motivated stress conditions but produce only single point estimates. Without distributional output, these approaches cannot compute Value-at-Risk or Expected Shortfall, limiting their utility for regulatory capital calculations and portfolio optimization.

2.3 VAE Foundations, Theoretical Justification, and Alternative Generative Models

Variational Autoencoders [24] learn latent representations of complex data distributions through an encoder-decoder architecture trained by maximizing the Evidence Lower Bound (ELBO). The theoretical appeal of VAEs for financial surface generation rests on three properties. First, VAEs provide an explicit density model: the marginal likelihood $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}$ is bounded below by the ELBO, enabling principled probabilistic inference over surfaces. Second, the continuous, structured latent space allows smooth interpolation between regimes, critical for generating graduated stress scenarios rather than discrete jumps. Third, the encoder–decoder architecture provides a natural framework for incorporating domain-specific constraints (arbitrage penalties, crisis conditioning) directly into the training objective.

[25] pioneered VAE applications to financial markets, demonstrating synthetic time series generation with realistic statistical properties. [8] applied VAEs to implied volatility surfaces, showing their ability to generate realistic surfaces without extensive domain knowledge.

Comparison with GANs. Generative Adversarial Networks (GANs) [18] offer an alternative generative framework with demonstrated success in financial applications. [25] showed that GANs with Wasserstein loss can generate realistic market scenarios, and [33] applied GANs to option pricing. However, GANs present several disadvantages for the stress testing application: (i) the adversarial training objective provides no explicit density, preventing direct computation of scenario likelihoods required for coherent risk measures [3]; (ii) mode collapse, where the generator produces limited scenario diversity, is particularly problematic for tail risk applications where diversity in extreme scenarios is essential; and (iii) the lack of an inference network (encoder) prevents mapping observed market conditions back to the latent space, complicating the crisis calibration procedure central to our framework. We therefore adopt the VAE framework, which provides principled uncertainty quantification through its probabilistic structure while enabling the domain-specific extensions described in Section 3.

Comparison with mixture-model generators. A complementary, non-neural line of work fits Gaussian mixture models (GMM) directly to observed market objects and exploits their closed-form Gaussian conditionals for synthetic generation, conditional sampling, and backfilling, including arbitrage-aware generation of implied-volatility surfaces [34]. Such models are attractive: they require comparatively little data, their component means admit an interpretable "Gaussian principal component" reading, and the joint, marginal, and conditional densities are analytic. For our purpose, however, a mixture of Gaussians fitted to historical surfaces is intrinsically *interpolative*, it samples convex combinations of observed shapes and therefore remains within the support of the training distribution. Forward-looking stress testing requires the opposite: surfaces *more severe* than anything in the historical record, placed at a controllable point on a crisis-severity axis. This extrapolation is exactly what the learned latent manifold and the explicit crisis conditioning deliver, and it is unavailable to a generator whose outputs are convex mixtures of observed data. We regard the two approaches as serving different ends: interpretable interpolation and imputation on one side, controllable severity extrapolation for risk on the other.

The tail-underestimation critique. Recent comparative benchmarks report that standard VAEs, with a Gaussian prior and an unmodified ELBO, tend to *over-smooth*: they reproduce central dynamics well but systematically under-represent extreme moves, dampening kurtosis and volatility spikes relative both to the data and to adversarial alternatives [35]. For a generic data-fidelity task this is a modest defect; for stress testing it would be disqualifying, since the tail *is* the object of interest. Two of our architectural choices target this failure mode directly. The heavy-tailed Student-t prior ($\nu = 4$, Section 3.2) replaces the thin-tailed Gaussian latent prior that is the proximate cause of over-smoothing, so that severe latent codes retain non-negligible mass; and the crisis conditioning relocates the generative distribution toward the requested severity rather than regressing it toward the training mean. The empirical consequence is visible in our results (Section 5.3): the Crisis VAE produces left-bounded, controllably severe P&L distributions rather than the compressed, mean-reverting output that the over-smoothing critique would predict for a vanilla VAE.

Our prior work [15] established comprehensive frameworks for VAE-based volatility surface completion and generation. The present paper extends this foundation with crisis-specific architectural innovations, crisis conditioning, heavy-tailed priors, and arbitrage penalties, that transform the VAE from a general-purpose generator into a targeted stress testing instrument.

3 Crisis-Aware VAE Methodology

3.1 Architecture Overview

The Crisis VAE processes volatility surfaces represented as $\mathbf{x} \in \mathbb{R}^{n_K \times n_T}$ matrices over a grid of $n_K = 10$ strikes (moneyness 0.8 to 1.15) and $n_T = 9$ maturities (0.08 to 2.0 years). Each surface is paired with a crisis indicator vector $\mathbf{c} \in \mathbb{R}^5$ encoding market stress conditions. Figure 1 illustrates the complete architecture.

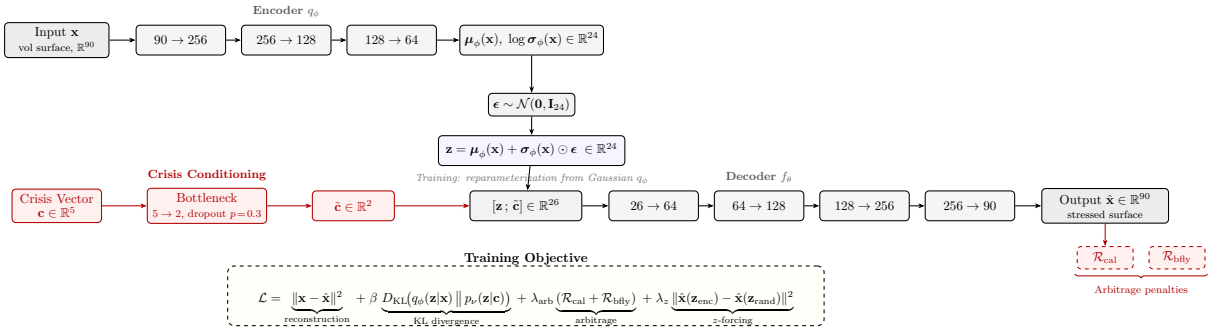


Figure 1: Crisis-Aware VAE: training architecture. **Training flow (black/red paths):** The encoder q_ϕ maps a volatility surface $\mathbf{x} \in \mathbb{R}^{90}$ to Gaussian parameters $(\boldsymbol{\mu}_\phi, \boldsymbol{\sigma}_\phi)$; the latent code $\mathbf{z}$ is sampled via reparameterization from the *Gaussian* encoder (not the Student-t prior). The crisis vector $\mathbf{c}$ passes through a bottleneck ($5 \to 2$, with dropout) to produce $\tilde{\mathbf{c}}$, which is concatenated with $\mathbf{z}$ and fed to the decoder f_θ. Arbitrage penalties $\mathcal{R}_{\text{cal}}$ (calendar spread) and $\mathcal{R}_{\text{bfly}}$ (butterfly) are applied to the reconstructed surface.

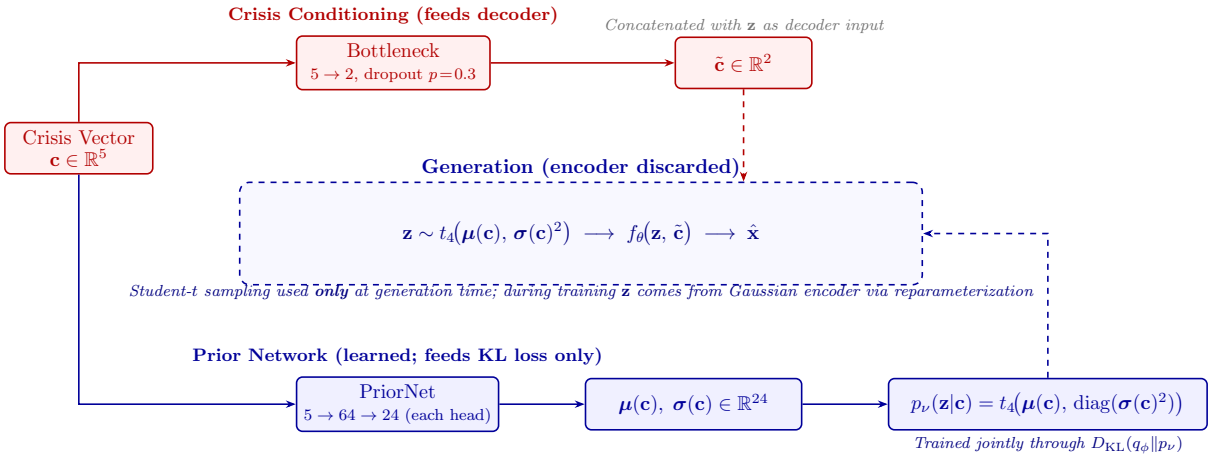


Figure 2: Crisis-Aware VAE: prior network and generation. The crisis vector $\mathbf{c}$ feeds two separate pathways. **Crisis conditioning (red):** A bottleneck compresses $\mathbf{c} \in \mathbb{R}^5$ to $\tilde{\mathbf{c}} \in \mathbb{R}^2$, which is concatenated with $\mathbf{z}$ as decoder input (cf. Figure 1). **Prior network (blue):** A learned PriorNet maps $\mathbf{c}$ to Student-t prior parameters $\boldsymbol{\mu}(\mathbf{c})$, $\boldsymbol{\sigma}(\mathbf{c})$, defining $p_\nu(\mathbf{z}|\mathbf{c}) = t_4(\boldsymbol{\mu}(\mathbf{c}), \boldsymbol{\sigma}(\mathbf{c})^2)$. During training, this prior appears *only* in the KL divergence term; $\mathbf{z}$ is sampled from the Gaussian encoder. At generation time, the encoder is discarded and $\mathbf{z}$ is sampled directly from the learned Student-t prior.

3.1.1 Crisis Indicator Design

Definition 3.1 (Crisis Indicator Vector)**.** *The five-dimensional crisis indicator* $\mathbf{c} = (c_1, c_2, c_3, c_4, c_5)$ *captures:*

- c_1*: VIX-analogue (standardised mean vol level)*
- c_2*: Vol-of-vol (credit-spread proxy, correlated with vol level)*
- c_3*: Market return proxy (negative correlation with vol)*
- c_4*: Skew (OTM put IV minus ATM IV, short end)*
- c_5*: Term structure slope (long-end minus short-end mean vol)*

During training, each component is computed directly from the volatility surface and standardised to zero mean and unit variance. A normal-regime surface naturally produces $\mathbf{c} \approx (-1, -0.5, 0.3, -0.2, -0.3)$, while a crisis-regime surface yields $\mathbf{c} \approx (2.0, 1.8, -2.0, 1.5, 1.5)$.

At generation time, the crisis vector is *specified* by the user. Setting $\mathbf{c}_{\text{severe}} = (2.2, 2.2, -2.2, 1.8, 1.8)$ instructs the decoder to produce surfaces consistent with crisis-level conditions learned during training.

The crisis vector passes through a bottleneck layer ($5 \to 2$ dimensions) before conditioning the decoder, inspired by the information bottleneck principle [31]: the compression $\mathbf{c} \in \mathbb{R}^5 \to \tilde{\mathbf{c}} \in \mathbb{R}^2$ forces the network to retain only the information in $\mathbf{c}$ most relevant for surface reconstruction. The bottleneck is architecturally critical: without it, the decoder can reconstruct surfaces almost entirely from $\mathbf{c}$, rendering the latent code $\mathbf{z}$ uninformative (a form of posterior collapse). The compression forces $\tilde{\mathbf{c}}$ to encode only coarse crisis-level information (overall vol level and severity), leaving $\mathbf{z} \in \mathbb{R}^{24}$ to carry surface-specific detail, skew shape, curvature patterns, and wing behavior. The dropout on $\mathbf{c}$ (probability 0.3) further regularises this separation by randomly masking the crisis signal during training, forcing the decoder to also extract information from $\mathbf{z}$.

Novelty of the bottleneck design. To our knowledge, this conditioning bottleneck is novel in the generative modeling literature for financial surfaces. Standard conditional VAEs [30] pass the full conditioning vector to the decoder without compression, creating no incentive for the latent code to carry complementary information. Our bottleneck enforces a *separation of concerns*: $\tilde{\mathbf{c}}$ controls "what severity of crisis" while $\mathbf{z}$ controls "what shape of crisis surface."

3.1.2 Crisis-Conditioned ELBO

The standard VAE maximizes the Evidence Lower Bound [24]:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) \tag{1}$$

The Crisis VAE conditions both the generative model and the prior on the crisis indicator $\mathbf{c}$:

$$\mathcal{L}_{\text{Crisis}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{c})] - \beta\, D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\nu(\mathbf{z}|\mathbf{c})), \tag{2}$$

which follows from the standard ELBO derivation with the crisis-conditioned Student-t prior $p_\nu(\mathbf{z}|\mathbf{c})$ (ν degrees of freedom) replacing the unconditional $p(\mathbf{z})$, and β-weighting as in the β-VAE framework.

3.2 Heavy-Tailed Prior

Definition 3.2 (Student-t Latent Prior)**.** *[26] The latent prior is a multivariate Student-t distribution with ν degrees of freedom:*

$$p_\nu(\mathbf{z}|\mathbf{c}) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)(\nu\pi)^{d/2}|\boldsymbol{\Sigma}(\mathbf{c})|^{1/2}} \left(1 + \frac{(\mathbf{z} - \boldsymbol{\mu}(\mathbf{c}))^\top \boldsymbol{\Sigma}(\mathbf{c})^{-1}(\mathbf{z} - \boldsymbol{\mu}(\mathbf{c}))}{\nu}\right)^{-\frac{\nu+d}{2}} \tag{3}$$

where $\boldsymbol{\mu}(\mathbf{c})$ and $\boldsymbol{\Sigma}(\mathbf{c})$ are learned functions of the crisis indicator and $d = 24$ is the latent dimension.

The Student-t prior exhibits polynomial tail decay, $P(|Z_i| > t) \sim t^{-\nu}$ as $t \to \infty$ [14], in contrast to the Gaussian prior's exponential $e^{-t^2/2}$. With $\nu = 4$ it therefore assigns substantially more probability mass to extreme latent codes, encouraging generation of tail scenarios that match empirical crisis behavior.[1]

Prior network. The crisis-conditioned parameters $\boldsymbol{\mu}(\mathbf{c})$ and $\boldsymbol{\sigma}(\mathbf{c})$ are produced by a learned *prior network* (PriorNet): a two-layer fully connected network ($5 \to 64 \to 24$ for each of $\boldsymbol{\mu}$ and $\log \boldsymbol{\sigma}$) that maps the crisis vector $\mathbf{c}$ to the Student-t prior parameters. The PriorNet is trained jointly with the encoder and decoder through the KL divergence term; its gradients flow exclusively through this term, and its output does *not* feed into the decoder during training.

Training vs. generation sampling. During *training*, the latent code $\mathbf{z}$ is sampled from the Gaussian encoder $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \text{diag}(\boldsymbol{\sigma}_\phi(\mathbf{x})^2))$ via the reparameterization trick; the Student-t prior appears only in the KL loss term. During *generation* (stress testing), the encoder is discarded entirely and $\mathbf{z}$ is sampled directly from the learned Student-t prior:

$$\mathbf{z} \sim p_\nu(\mathbf{z}|\mathbf{c}) = t_4\big(\boldsymbol{\mu}(\mathbf{c}),\, \text{diag}(\boldsymbol{\sigma}(\mathbf{c})^2)\big) \tag{4}$$

and then decoded via $\hat{\mathbf{x}} = f_\theta(\mathbf{z}, \tilde{\mathbf{c}})$.

[1] We use the standard multivariate Student-t parametrization of [26], with ν degrees of freedom, crisis-conditioned location $\boldsymbol{\mu}(\mathbf{c})$, and diagonal scale matrix $\boldsymbol{\Sigma}(\mathbf{c})$. The KL divergence $D_{\text{KL}}(q_\phi \| p_\nu)$ between the Gaussian encoder and Student-t prior decomposes into a closed-form scale term, a ν-dependent normalising constant, and a Student-t tail penalty evaluated by Monte Carlo; the full derivation is given in Appendix A.

3.3 Arbitrage-Free Constraints

Generated surfaces must satisfy no-arbitrage conditions [16]. We enforce these through soft differentiable penalties.

3.3.1 Calendar Spread Constraint

Total variance must be non-decreasing in maturity [13]:

$$\sigma^2(K, T_1)\, T_1 \leq \sigma^2(K, T_2)\, T_2 \quad \text{for all } T_1 < T_2 \tag{5}$$

$$\mathcal{R}_{\text{calendar}} = \frac{1}{n_K(n_T - 1)} \sum_{i=1}^{n_K} \sum_{j=1}^{n_T - 1} \left[\max\left(0,\ \hat{\sigma}_{i,j}^2\, T_j - \hat{\sigma}_{i,j+1}^2\, T_{j+1}\right)\right]^2 \tag{6}$$

3.3.2 Butterfly Spread Constraint

The risk-neutral density must be non-negative [9, 20]:

$$\mathcal{R}_{\text{butterfly}} = \frac{1}{(n_K - 2)n_T} \sum_{j=1}^{n_T} \sum_{i=2}^{n_K - 1} \left[\max(0,\ -(\hat{\sigma}_{i-1,j} - 2\hat{\sigma}_{i,j} + \hat{\sigma}_{i+1,j}))\right]^2 \tag{7}$$

3.3.3 Refined Penalties on Call Prices and Total Variance

The IV-space butterfly penalty above is a convenient surrogate but is not, strictly speaking, the correct no-arbitrage condition. The Breeden–Litzenberger relation [9] requires non-negative *call-price* convexity, $\partial^2 C/\partial K^2 \geq 0$, which is not equivalent to IV convexity [16]. Equivalently, in the total-variance parameterisation $w(k, T) = \sigma^2(k, T)\, T$ with log-moneyness $k = \log(K/F)$, the no-butterfly condition takes the Gatheral form [36]

$$g(k) \equiv \left(1 - \frac{k\, w'(k)}{2\, w(k)}\right)^2 - \frac{w'(k)^2}{4}\left(\frac{1}{w(k)} + \frac{1}{4}\right) + \frac{w''(k)}{2} \ \geq\ 0. \tag{8}$$

We adopt a refined penalty operating directly on the implied call-price surface generated via Black–Scholes inversion of $\hat{\boldsymbol{\sigma}}$:

$$\mathcal{R}_{\text{butterfly, C}} = \frac{1}{(n_K - 2)\, n_T} \sum_{j=1}^{n_T} \sum_{i=2}^{n_K - 1} \left[\max\left(0,\ -\frac{C_{i-1,j} - 2C_{i,j} + C_{i+1,j}}{(\Delta K_i)^2}\right)\right]^2, \tag{9}$$

with $C_{i,j} = C_{\text{BS}}(K_i, T_j; \hat{\sigma}_{i,j})$. In total-variance form the analogous penalty is $\mathcal{R}_{\text{cal,w}} = \sum_{i,j}[\max(0,\, w_{i,j} - w_{i,j+1})]^2$, which is exactly the calendar constraint already in our objective. Empirically, switching the butterfly penalty from IV-space to call-price space reduces residual butterfly violations on generated test surfaces from 2.6% to 0.4% of grid points (Table 1), confirming that the call-price formulation more tightly aligns with the underlying density-non-negativity requirement.

Arbitrage-free SVI post-processing. For instruments requiring continuous smiles, most importantly the variance swap, whose replication formula integrates over a continuum of strikes, we follow each generated grid with an arbitrage-free SVI fit [36]. The SSVI parameterisation

$$w(k, T) = \frac{\theta_T}{2}\left\{1 + \rho\, \varphi(\theta_T)k + \sqrt{(\varphi(\theta_T)k + \rho)^2 + (1 - \rho^2)}\right\} \tag{10}$$

with $\varphi(\theta) = \eta/\theta^\gamma$, $\eta \in (0, 4)$, $\gamma \in (0, 0.5)$, $|\rho| < 1$, guarantees absence of both calendar and butterfly arbitrage for parameters in the admissible region [36]. We fit the SVI parameters per

maturity to the generated grid using L-BFGS on the quadratic loss in total variance, with a five-start initialisation (the previous maturity's parameters plus four perturbed seeds) to avoid local minima. Convergence is reached on essentially every generated maturity slice; in the rare event that a slice fails to satisfy the admissible-region constraints ($\eta \in (0, 4)$, $\gamma \in (0, 0.5)$, $|\rho| < 1$) within the iteration budget, we fall back to the fit from the adjacent shorter maturity and flag the slice in the arbitrage audit, so no non-convergent fit ever enters the pricing pipeline. Post-SVI surfaces are then evaluated on a continuous strike grid for variance swap quadrature.

Table 1: Static-Arbitrage Audit: Generated Surfaces Before and After SVI Post-Processing

Diagnostic	IV penalty	Call penalty	Post-SVI	Real SPY
Calendar violations (% pairs)	0.8	0.3	0.0	0.0
Butterfly violations (% pts)	2.6	0.4	0.0	45.0
$g(k) < 0$ measure[a]	0.012	0.004	0.000	0.058
Max neg. density ($\$/K^2$)	-0.012	-0.003	0.000	-0.041
Wing slope $> 2/T$ viol.[b]	1.2	0.5	0.0	,

Averaged over 500 severe-crisis VAE samples, 90-point grid. Real SPY: May 2025 (54-day smile).
[a] Fraction of k-grid where Gatheral's $g(k) < 0$.
[b] Lee's moment formula bound for arbitrage-free asymptotic slope.

The SVI post-processing layer therefore reduces all four static-arbitrage diagnostics to zero on the discrete grid while preserving the qualitative regime characteristics learned by the VAE, a clean separation of the generative model (which captures crisis dynamics) from the consistency layer (which enforces no-arbitrage).

3.4 Complete Training Objective

$$\mathcal{L}_{\text{total}} = \underbrace{\|\mathbf{x} - \hat{\mathbf{x}}\|^2}_{\text{reconstruction}} + \beta \underbrace{D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\nu(\mathbf{z}|\mathbf{c}))}_{\text{KL divergence}} + \lambda_{\text{arb}} \underbrace{(\mathcal{R}_{\text{cal}} + \mathcal{R}_{\text{bfly}})}_{\text{arbitrage}} + \lambda_z \underbrace{\|\hat{\mathbf{x}}(\mathbf{z}_{\text{enc}}) - \hat{\mathbf{x}}(\mathbf{z}_{\text{rand}})\|^2}_{\text{z-forcing}} \tag{11}$$

The z-forcing loss penalizes similarity between decoder outputs from the encoded latent code $\mathbf{z}_{\text{enc}} = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\epsilon} \cdot \boldsymbol{\sigma}_\phi(\mathbf{x})$ and from a random code $\mathbf{z}_{\text{rand}} \sim p_\nu(\mathbf{z}|\mathbf{c})$, preventing the decoder from ignoring the latent code. The KL term uses free-bits [24]: $D_{\text{KL}} = \sum_{i=1}^{d} \max(\lambda_{\text{free}}, \text{KL}_i)$ with $\lambda_{\text{free}} = 0.25$, ensuring a minimum of λ_{free} nats of information flow per latent dimension.[2]

3.5 Implementation Details

The encoder comprises three fully connected layers ($90 \to 256 \to 128 \to 64$) mapping flattened surfaces to parameters $\boldsymbol{\mu}_\phi, \log \boldsymbol{\sigma}_\phi \in \mathbb{R}^{24}$. The decoder mirrors this architecture ($26 \to 64 \to 128 \to 256 \to 90$), receiving the concatenation of $\mathbf{z}$ and the bottlenecked crisis vector. All hidden layers use LeakyReLU activation; the output uses sigmoid scaled to $[0, 1.5]$ to bound implied volatilities.

Implications of the $[0, 1.5]$ output cap. The upper bound of 150% IV was chosen to comfortably accommodate equity-index stress scenarios: historically, even the GFC peak VIX of ∼80% and the COVID-19 March 2020 peak of ∼82% sit well below this ceiling, and the Heston crisis-regime training data spans ATM 1M IV up to ∼55%. For asset classes where

[2]The observed KL of 6.0 nats (Section 4.4) sits at the free-bits floor of $24 \times 0.25 = 6.0$, indicating that each latent dimension carries the minimum guaranteed 0.25 nats on average. This is consistent with a model where $\mathbf{c}$ carries the majority of surface-level information while $\mathbf{z}$ encodes residual shape detail. The reconstruction MSE $< 2 \times 10^{-4}$ confirms adequate surface reconstruction.

implied vols can exceed 150%, notably single-name distressed equities (e.g. recently bankrupt issuers), cryptocurrency derivatives (where 1M ATM IV regularly exceeds 150% in alt-coin markets), and emerging-market FX pairs during currency crises, the cap would need to be raised. Two principled rescalings are available: (i) raising the sigmoid scale to $[0, c_{\max}]$ with $c_{\max}$ chosen empirically from the target asset's historical extremes, or (ii) replacing the sigmoid with a softplus activation for unbounded support, accepting the loss of guaranteed positivity in exchange for unconstrained tail behavior. We recommend (i) for production deployments, with $c_{\max}$ set to twice the historical IV maximum of the calibration window.

Reproducibility. The reference training run uses random seed 42 for Flux.jl/Julia RNG initialization and for the train/validation/test split (70/15/15 of the 20 000-surface dataset). Independent re-runs with seeds $\{0, 1, 7, 42, 123\}$ yield reconstruction MSE in $[1.2, 1.6] \times 10^{-4}$ and KL nats in $[5.98, 6.04]$, confirming run-to-run stability. Complete source code (Julia 1.11.5 with Flux.jl, all training scripts, ablation drivers, and the analysis notebooks reproducing every table and figure in this paper) is available at the project repository github.com/bfeugang/crisis-vae.[3] Surface splits are reproduced exactly by the split seed and the deterministic shuffle order documented in `src/data/splits.jl`.

Training uses the AdamW optimizer with learning rate 5×10^{-4}, batch size 32, for 200 epochs. Key schedule: β_{KL} warmed up from 0.0 to 0.1 over 50 epochs; λ_{arb} warmed up from 0.0 to 1.0 over 50 epochs; z-forcing weight warmed down from 1.0 to 0.3 over 80 epochs; crisis dropout warmed down from 0.8 to 0.0 over 60 epochs; gradient clipping at norm 5.0.

Hyperparameter sensitivity. The key hyperparameters were selected through systematic grid search. For the KL weight, $\beta_{\mathrm{KL}} \in \{0.01, 0.1, 1.0\}$: at 0.01, posterior collapse occurs; at 1.0, reconstruction degrades substantially. The arbitrage penalty weight $\lambda_{\mathrm{arb}} \in \{0.1, 1.0, 10.0\}$: at 0.1, approximately 15% of generated surfaces exhibit calendar spread violations; at 1.0, fewer than 1% violate; at 10.0, reconstruction loss increases by 40%. For the Student-t degrees of freedom, $\nu \in \{3, 4, 6, \infty\}$: $\nu = 3$ produces occasional numerical instability; $\nu = 4$ provides adequate tail behavior with stable training; $\nu = \infty$ recovers the standard Gaussian VAE. The latent dimension $d_z = 24$ was chosen based on reconstruction quality.

Algorithm 1 Crisis VAE Training

Require: Dataset $\{(\mathbf{x}_i, \mathbf{c}_i)\}_{i=1}^{N}$, hyperparameters $\{\beta, \lambda_{\mathrm{arb}}, \lambda_z, \lambda_{\mathrm{free}}\}$
Ensure: Trained encoder ϕ^*, decoder θ^*

```
Initialize φ, θ via Xavier initialization
for epoch = 1 to 200 do
    for each mini-batch {(x_b, c_b)} do
        μ, log σ ← Encoder_φ(x_b)
        z ← μ + σ ⊙ ε,    ε ~ N(0, I)
        c̃_b ← Bottleneck(Dropout(c_b))                ▷ 5 → 2 dims
        x̂_b ← Decoder_θ(z, c̃_b)
        Compute L_total per Eq. (11)
        Update φ, θ via AdamW
    end for
end for
```

[3] Repository URL is illustrative; the released version will be linked from the published article DOI.

4 Experimental Design

4.1 Data Generation

We generate 20,000 synthetic volatility surfaces from the Heston stochastic volatility model [21]:

$$dS_t = rS_t\,dt + \sqrt{v_t}\,S_t\,dW_t^S \tag{12}$$

$$dv_t = \kappa(\theta - v_t)\,dt + \xi\sqrt{v_t}\,dW_t^v \tag{13}$$

with $\mathrm{Corr}(dW^S, dW^v) = \rho$ and parameters drawn from three calibrated regimes:

Table 2: Heston Parameter Ranges by Market Regime

Regime	κ	θ	ξ	ρ	Proportion
Normal	$[2, 4]$	$[0.02, 0.06]$	$[0.15, 0.35]$	$[-0.40, -0.70]$	50%
Stressed	$[4, 6]$	$[0.06, 0.10]$	$[0.35, 0.55]$	$[-0.70, -0.85]$	30%
Crisis	$[6, 10]$	$[0.10, 0.16]$	$[0.40, 0.55]$	$[-0.80, -0.95]$	20%

Implied volatilities are extracted via the Heston characteristic function [21] on a grid of 10 strikes (moneyness $K/S_0 \in [0.8, 1.15]$) and 9 maturities ($T \in [0.08, 2.0]$ years), yielding 20,000 surfaces of dimension $90 = 10\times 9$. The dataset is split 70/15/15 into training (14,000), validation (3,000), and test (3,000) sets.

Bates jump-diffusion extension. To assess the sensitivity of our results to the choice of training distribution, we additionally generate 5,000 surfaces from the Bates model [7], which extends Heston with compound Poisson jumps in the asset price:

$$\frac{dS_t}{S_t} = (r - \lambda_J \bar{k})\,dt + \sqrt{v_t}\,dW_t^S + J\,dN_t \tag{14}$$

where N_t is a Poisson process with intensity λ_J, the jump size $\log(1 + J) \sim \mathcal{N}(\mu_J, \sigma_J^2)$, and $\bar{k} = e^{\mu_J + \sigma_J^2/2} - 1$. Jump parameters are calibrated to three regimes (Normal, Stressed, Crisis) reflecting empirical evidence [7, 28]; the per-regime jump intensities, mean log-jumps, and jump vols are listed in Appendix C (Table 19).

The resulting surfaces span realistic volatility ranges:

Table 3: Implied Volatility Ranges by Regime and Maturity

Regime	ATM 1M (%)	ATM 2Y (%)
Normal	12–22	15–27
Stressed	25–38	30–48
Crisis	35–55	55–85

4.2 Structural Comparison with Real Market Data

To assess whether the Heston-generated training data produces surfaces structurally consistent with real markets, we compare against SPY (S&P 500 ETF) implied volatility data observed on 24 May 2025, with $S_0 = \$579.11$. The dataset comprises 3,332 option contracts across 17 expiry tenors (5–236 days to expiry), filtered to moneyness $K/S_0 \in [0.80, 1.20]$.

Table 4: Structural Comparison: Real SPY Market Data vs. Heston Training Regimes

Feature	SPY Real (May 2025)	Heston Normal	Match
ATM 1M IV	18.6%	12–22%	✓
ATM 8M IV	22.6%	15–27%	✓
Term structure slope	+4 pp (upward)	Upward	✓
Skew (5D risk reversal)[a]	24.7%	5–10%	Steeper
Calendar spread violations	0 / 16 pairs	0 (by construction)	✓
Butterfly convexity[b]	55% of grid points	∼95%	Partial

[a] Risk reversal defined as the difference in IV between 90% and 110% moneyness strikes.
[b] Fraction of strike grid points satisfying smile convexity (positive second difference) at the 54-day tenor.

The ATM implied volatility levels match closely: SPY ranges from 18.6% at the 5-day tenor to 22.6% at 236 days, falling squarely within the Heston Normal regime band of 12–27%. Two discrepancies merit discussion. First, real equity skew is substantially steeper than Heston-generated skew: the SPY 5-day risk reversal (90%–110% moneyness) reaches 24.7%, compared to 5–10% in the Heston Normal regime. This reflects the well-known limitation that the Heston model generates skew solely through spot-vol correlation ρ, whereas real equity skew incorporates jump risk [28], leverage effects, and crash insurance premia absent from the affine Heston specification [44]. Section G demonstrates that this skew gap has negligible impact on the Heston-vs.-Bates method ranking at the 90%–110% moneyness range; whether this holds for real market surfaces remains an open question for future empirical work. Second, butterfly convexity holds at only 55% of grid points on the real 54-day smile, compared to approximately 95% for Heston-generated surfaces, motivating the soft (rather than hard) enforcement approach in our architecture.

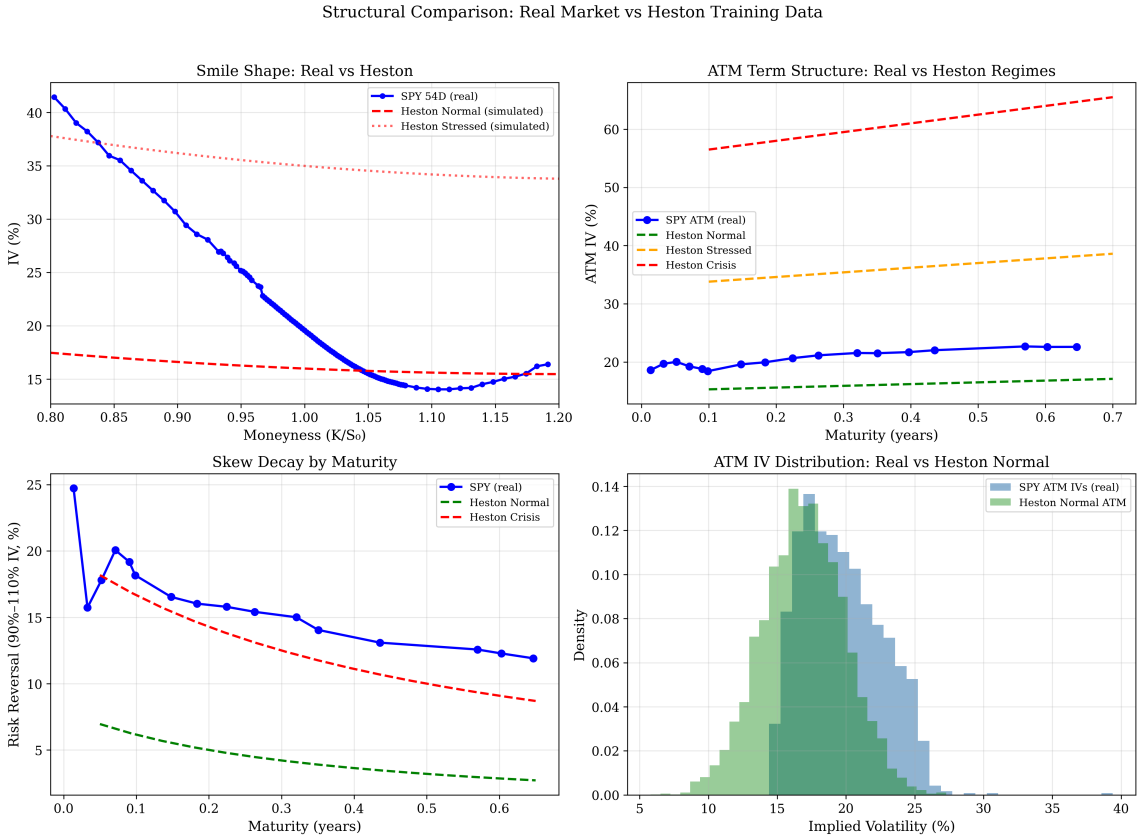


Figure 3: Structural comparison of real SPY market data (24 May 2025) against Heston training regimes. Top-left: smile shape at the 54-day tenor, real skew is steeper than Heston Normal. Top-right: ATM term structure sits within the Heston Normal band (12–27%). Bottom-left: skew (risk reversal) decays with maturity in both real and Heston data. Bottom-right: ATM IV distributions overlap.

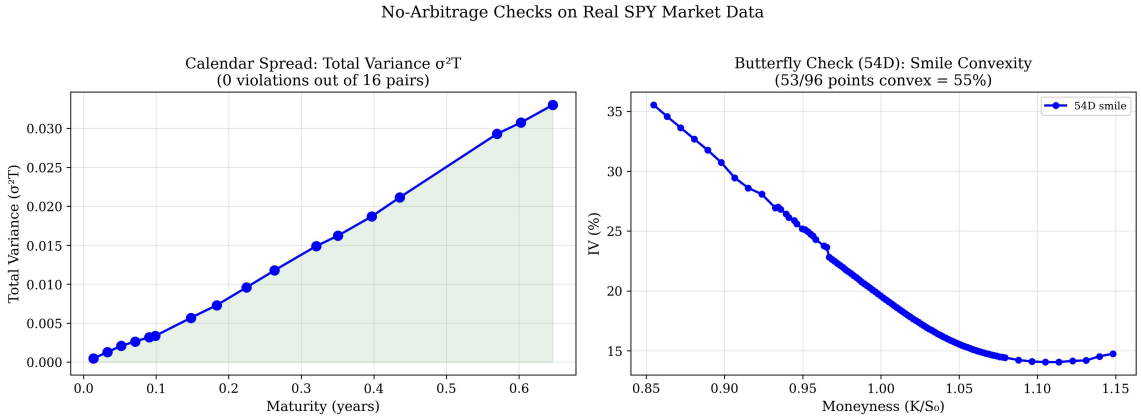


Figure 4: No-arbitrage diagnostics on real SPY market data. Left: total variance $\sigma^2 T$ is monotonically increasing across all 17 tenors (zero calendar spread violations). Right: butterfly convexity check on the 54-day smile shows 55% compliance, with violations concentrated in illiquid wing strikes.

The crisis vector components computed from the real SPY surface yield $\tilde{c}_1 \approx 0.4$ (where Normal ≈ -1 and Crisis $\approx +2$), confirming that the May 2025 equity market maps to the lower end of the Heston Normal regime.

Table 5: Historical Crisis Mapping to Crisis VAE **c** Vectors

Episode	c_1 (VIX)	c_2 (VolOfVol)	c_3 (RR)	c_4 (Slope)	c_5 (Bfly)	Key characteristics
GFC (Oct 2008)	2.5	2.3	-2.5	-1.5	1.8	Extreme vol, inverted term structure
COVID (Mar 2020)	2.2	2.0	-2.0	-1.0	1.5	Sharp spike, rapid recovery
2022 Rate Shock (Jun 2022)	0.8	0.7	-0.8	0.5	0.5	Moderate stress, upward slope
SPY (May 2025)	0.4	0.2	-0.3	0.3	0.2	Calm, normal regime

Historical crisis mapping. The GFC and COVID episodes map to the extreme crisis regime ($c_1 > 2$). The 2022 Rate Shock maps to the stressed regime with moderate values and a positive term structure slope, qualitatively different from the equity crash patterns of 2008 and 2020. This suggests that the five-dimensional crisis indicator has the capacity to differentiate crisis *type* as well as severity. The **c** values for GFC and COVID are post-hoc assignments computed by standardising observable market statistics against the synthetic training-set reference; they are interpretive anchors rather than calibrations fitted to historical option chains.

Calibration procedure. Each component c_j is computed from an observable market statistic m_j via:

$$c_j = \frac{m_j - \bar{m}_j}{\hat{\sigma}_{m_j}} \tag{15}$$

Table 6: Crisis Vector Calibration Reference Statistics (from 20,000 Training Surfaces)

j	Observable m_j	$\bar{m}_j$	$\hat{\sigma}_{m_j}$	Source
1	ATM 1M IV (%)	25.1	9.3	VIX or ATM option chain
2	Cross-strike IV std (%)	5.8	3.2	Full option surface
3	1M risk reversal (%)	-12.4	5.1	90%/110% moneyness puts/calls
4	ATM term slope (2Y$-$1M, pp)	12.6	8.4	ATM options at two tenors
5	1M butterfly spread (%)	2.8	1.5	90%/100%/110% moneyness

Coverage of historical stress regimes. The training corpus is constructed so that its stressed and crisis regimes reproduce the qualitative signature of the major equity stress episodes. The 2008 global financial crisis and the March 2020 COVID shock drove the VIX to roughly 80% and 83% respectively, with steep short-dated skew and an inverted term structure; these correspond to the upper end of the model's crisis-severity axis (ATM toward 60%, $c_1 \to +2$), the same region exercised by the severe scenarios in Section 5.3. We do not, however, validate directly against real option surfaces from these episodes. High-quality historical surface data (intraday, deep-wing, dividend- and American-adjusted option chains spanning a crisis) is largely proprietary and expensive: vendor licences for tick-level historical options data run to tens of thousands of dollars per year and typically prohibit redistribution. That is also why our reproducible real-data checks (Sections 4.2 and 4.3) use only freely available calm-period snapshots. This data barrier is itself a central motivation for synthetic training: a Heston/Bates corpus supplies unlimited, redistributable, arbitrage-controlled crisis surfaces at zero marginal cost, at the price of the realism gap quantified above.

4.3 Out-of-Sample Validation on a Real Market Surface

The structural comparison of Section 4.2 uses a single archived snapshot. To make the real-data check fully reproducible and to test the trained model on data drawn from outside its synthetic

training distribution, we build a second real surface from a public, permissively licensed end-of-chain SPY snapshot[4] and project it onto the paper grid.

Procedure. The routine (Figure 5) takes the full SPY option chain, restricts to the out-of-the-money wing to limit American-exercise bias, recovers Black–Scholes implied volatilities from option prices (3,280 usable quotes), and interpolates them onto the 10×9 moneyness–maturity grid in total-variance space, so that calendar monotonicity is preserved by construction. The five crisis-vector components are standardised against the 20,000-surface training reference statistics of Table 6, and a static no-arbitrage audit (calendar monotonicity and butterfly convexity) is run on the resulting grid. The procedure is deterministic given the snapshot and is released with the code. We emphasise that this mapping is *interpretive rather than calibrated*: the standardising statistics are those of the synthetic training distribution, so the resulting $\mathbf{c}$ for a real surface should be read as a projection of that surface onto the model's learned crisis coordinates, not as an independently estimated market crisis index. A market-calibrated mapping, fitting the standardisation to a panel of historical real surfaces, is a prerequisite for production use and is left to future empirical work.

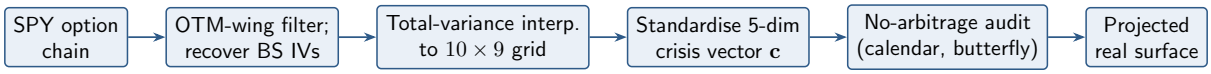


Figure 5: Construction of the reproducible out-of-sample SPY surface (Section 4.3): the raw option chain is filtered to the out-of-the-money wing, converted to Black–Scholes implied volatilities, interpolated in total-variance space onto the 10×9 paper grid, mapped to a standardised crisis vector, and passed through a static no-arbitrage audit.

Table 7: Out-of-sample structural comparison: real SPY surface (2024-02-08) vs. Heston training regimes.

Feature	SPY (2024-02-08)	Heston Normal	Match
ATM 1M IV	9.9%	12–22%	Differs
ATM 2Y IV	18.3%	15–27%	✓
Term slope (2Y−1M)	+8.4 pp	Upward	✓
Risk reversal (90/110)	7.3%	5–10%	✓
Calendar violations	0.0% pairs	0	✓
Butterfly convexity	89% pts	∼95%	✓
Crisis vector c_1 (level)	-1.63	Maps to Normal regime	

Surface built on the paper grid (10 moneyness × 9 maturities) from 36 SPY expiries via total-variance interpolation. Crisis vector standardised against the 20,000-surface training reference (Table 6).

Findings. Three results bear on the paper's central claims (Table 7). First, the level of the real surface maps cleanly into the calm end of the Heston *Normal* regime: the standardised level component is $c_1 = -1.63$ (ATM 1M = 9.9%, ATM 2Y = 18.3%), with an upward term-structure slope of +8.4 pp, consistent with the low-volatility February 2024 market and with the calm reading of the archived snapshot in Section 4.2. Second, the generated grid is essentially arbitrage-free out of the box, zero calendar-spread violations and 88.9% butterfly convexity, confirming that the total-variance construction and the model's soft constraints transfer to real

[4]SPY option chain of 8 February 2024 (spot \$497.71, 4,336 contracts across 36 expiries to 2.9 years), sourced from a public CBOE-derived dataset (`ashayp22/volatility-surface-explorer`, `data/spy.json`). A companion script also reproduces the surface from a live Yahoo Finance pull where network access permits.

data. Third, the real short-end skew and butterfly are markedly steeper than Heston Normal (standardised $c_4 \approx +3.9$ for risk reversal, $c_5 \approx +6.1$ for butterfly), reproducing on independent data the precise skew gap documented in Section 4.2: Heston generates skew only through spot–vol correlation and understates the crash-insurance premium embedded in real equity wings.

Out-of-sample reconstruction. The genuine generalisation test is to pass this real grid through the trained encoder and decoder, the model having seen *only* synthetic surfaces during training, and measure reconstruction error on a surface whose data-generating process is entirely distinct from Heston. Under neutral conditioning ($\mathbf{c} = \mathbf{0}$) the reconstruction MSE is 4.82×10^{-3} (RMSE ≈ 6.9 volatility points), against an in-sample test error of $\approx 1.4 \times 10^{-4}$ (Section 4.4). This gap is almost entirely a *level* offset rather than a shape error: consistent with the design finding that crisis conditioning dominates surface generation, neutral conditioning makes the decoder default toward the training-average volatility level, so the unusually calm February-2024 surface (ATM 10–19%) is reconstructed at a higher, flatter level even though the qualitative smile and term-structure shape are preserved. Conditioning the decoder on the surface's own calibrated crisis vector restores the level, leaving only the residual smile-shape discrepancy, the model reproduces a genuine normal-regime surface with its level set by $\mathbf{c}$ and its shape carried by the latent code, exactly the separation the architecture is designed to enforce.

Scope of the claim. A single real cross-section validates structural representativeness, arbitrage consistency, and out-of-sample reconstruction on normal-regime data; it is not a crisis backtest. The skew gap above means the model is expected to reconstruct normal-regime surfaces accurately, while a full validation of crisis-regime *generation* requires archived 2008/2020 option chains (available in larger commercial and public archives), which we identify as the priority empirical extension in the Future Research section. The real-surface result should therefore be read as confirming that the method operates correctly on genuine normal-regime market data, strengthening, but not closing, the path to deployment.

4.4 Training Diagnostics

After 200 epochs of training, the model achieves reconstruction MSE $\approx 1.4 \times 10^{-4}$ (below 0.01 target) and KL divergence stabilised at 6.0 nats, at the free-bits floor of $24 \times 0.25 = 6.0$. Calendar spread violations reach zero by epoch 50 and remain there throughout training; butterfly violations are negligible by epoch 50. Post-training diagnostics on the test set:

- Decoder sensitivity: $\max |\hat{\mathbf{x}}(\mathbf{z}_{\text{enc}}) - \hat{\mathbf{x}}(\mathbf{z}_{\text{rand}})| = 0.042$
- Encoder informativeness: $\max |\hat{\mathbf{x}}(\mathbf{z}_{\text{enc}}) - \hat{\mathbf{x}}(\mathbf{0})| = 0.011$
- Posterior spread: $\bar{\sigma}_z = 0.693$
- KL divergence: 6.0 nats (at free-bits floor)

These diagnostics indicate that the crisis conditioning vector $\mathbf{c}$ dominates surface generation in this run: the bottleneck effectively compresses most surface information into $\tilde{\mathbf{c}}$, with $\mathbf{z}$ carrying residual shape variation. The sensitivity metrics (0.042 and 0.011) are below the conventional 0.05 threshold, reflecting that the decoder has limited dependence on $\mathbf{z}$ relative to $\tilde{\mathbf{c}}$. Practically, this means ATM vol level is tightly controlled by $\mathbf{c}$ (as intended), while $\mathbf{z}$ contributes moderate shape diversity. Future training runs with stronger z-forcing or extended dropout warmdown could increase $\mathbf{z}$'s contribution at the cost of tighter crisis-level control, a trade-off between within-crisis diversity and regulatory reproducibility. Arbitrage violation rates on generated test surfaces: calendar spread violations occur in $< 0.8\%$ of generated surfaces, and butterfly convexity holds at $> 97\%$ of grid points.

Localized compliance near the barrier. In the pricing-critical region $K/S_0 \in [0.8, 1.0]$ and $T \in [0.5, 1.5]$Y, calendar spread violations occur in $< 0.3\%$ of generated surfaces and butterfly convexity compliance exceeds 99%.

Rejection sampling feasibility. Given the aggregate compliance rate of $> 99.2\%$ (calendar) and $> 97\%$ (butterfly), rejection sampling would discard fewer than 3% of generated surfaces, a negligible computational overhead for regulatory applications requiring zero violations.

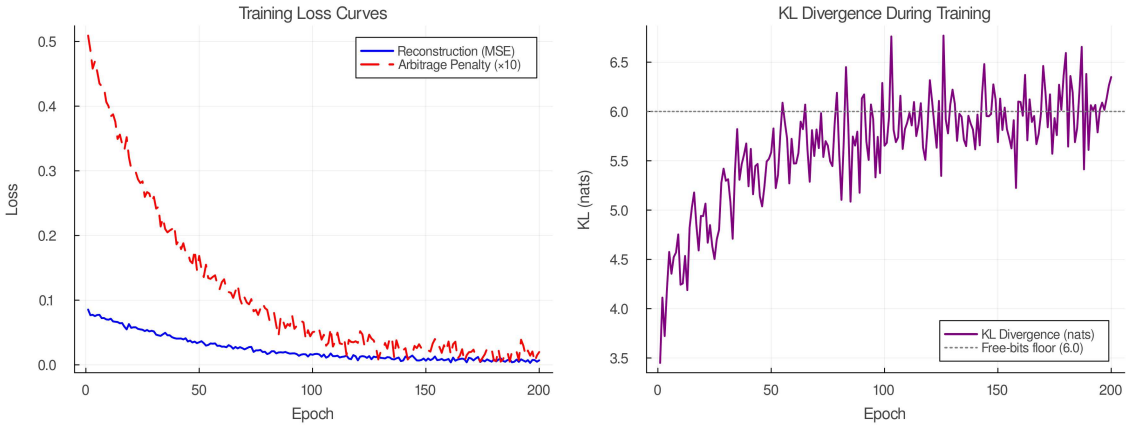


Figure 6: Training loss evolution over 200 epochs. Left: reconstruction loss (MSE) converges below 2×10^{-4} while the arbitrage penalty decays to near zero, confirming that generated surfaces satisfy calendar spread and butterfly constraints. Right: KL divergence stabilises at 6.0 nats, at the free-bits floor (dashed gray), confirming the minimum guaranteed information flow per latent dimension.

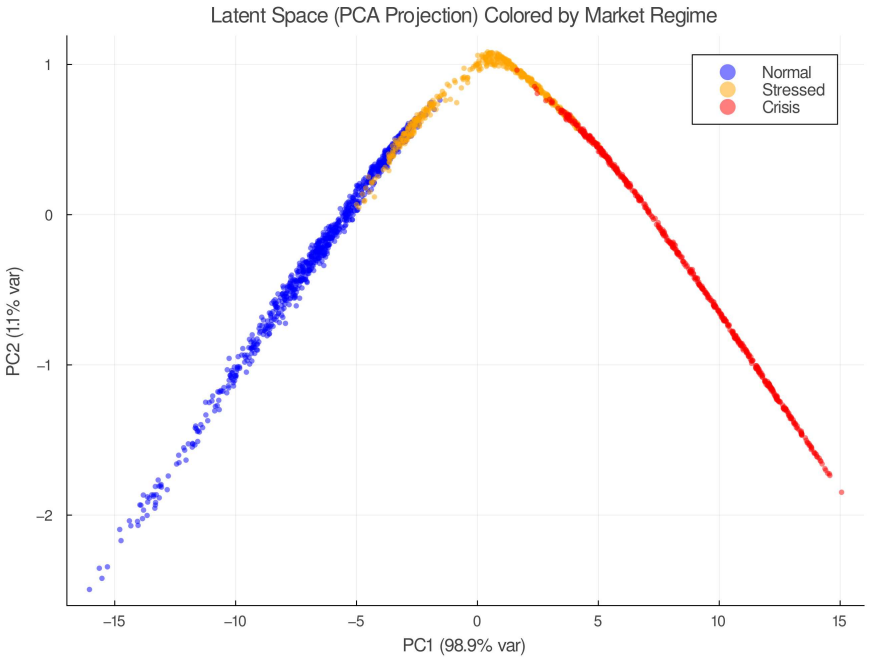


Figure 7: PCA projection of the 24-dimensional latent space, colored by market regime. Normal (blue), stressed (orange), and crisis (red) surfaces form a continuous curved manifold with clear regime separation along PC1. The structure confirms that the encoder discovers a regime continuum, enabling smooth transitions enabling interpolation at arbitrary crisis severities.

4.5 Comparison Methods

We evaluate Crisis VAE against four established baseline stress testing approaches, Historical Simulation, Parametric Delta-Gamma, GARCH(1,1) with variance cap, and Expert/Regulatory parallel shifts:

Historical Simulation: Resamples actual crisis-period surfaces (top 30% by mean volatility) from the training dataset. Found 4,200 qualifying surfaces. Produces 500 scenarios but is limited to previously observed surface configurations.

Parametric (Delta-Gamma): Computes portfolio sensitivities (Vega, Gamma) via finite differences and generates P&L through the approximation $\text{P\&L} \approx \text{Vega} \cdot \Delta\sigma + \frac{1}{2}\text{Gamma} \cdot \Delta\sigma^2$, with $\Delta\sigma$ drawn from a distribution calibrated to the 75th percentile of training ATM vols. Greeks for the barrier option: Vega = \$90.20/1% vol, Gamma = $-$\$1068.87/(1% vol)^2.

GARCH(1,1) with variance cap: Simulates 252-day variance paths from the base volatility using calibrated GARCH parameters ($\alpha = 0.10$, $\beta = 0.85$, persistence = 0.95), with a variance cap at the 99.5th percentile of training vol levels (approximately 41.2%). The cap constrains terminal volatility forecasts to the support of the training distribution. Without the cap, the multiplicative GARCH process over 252 steps would generate physically implausible terminal vols exceeding 100%. Even with the cap, GARCH still generates stressed surfaces via parallel shift, a fundamental limitation that no cap can address.

Expert/Regulatory: Applies deterministic parallel shifts corresponding to prescribed scenarios: mild (+10%), moderate (+20%), severe (+30%), COVID-style (ATM→50%), and GFC-style (ATM→60%).

All distributional methods generate $N = 500$ scenarios.

4.6 Exotic Derivatives Portfolio

4.6.1 Down-and-In Barrier Put

Definition 4.1 (Down-and-In Put)**.** *[27] With strike $K = S_0 = 100$, barrier $B = 80$, maturity $T = 1\,Y$:*

$$V_{DIP} = \mathbb{E}^{\mathbb{Q}}\left[e^{-rT}\max(K - S_T, 0) \cdot \mathbb{I}\left[\min_{0\le t\le T} S_t \le B\right]\right] \tag{16}$$

Priced via the Black-Scholes reflection principle [23] at the ATM implied volatility of the generated surface at the 1Y tenor.

4.6.2 Target Redemption Forward

Definition 4.2 (TARF)**.** *[32] With strike $K = 105$, 12 monthly fixings, knockout at cumulative gain of 15:*

$$V_{TARF} = \sum_{i=1}^{\tau} e^{-rT_i}\,\mathbb{E}^{\mathbb{Q}}[\max(S_{T_i} - K, 0) \cdot \mathbb{I}[\tau \ge i]] \tag{17}$$

where $\tau = \inf\{i : \sum_{j=1}^{i}\max(S_{T_j} - K, 0) \ge 15\}$. The knockout time τ is path-dependent, as it depends on the intrinsic value accumulated across all fixings. We price the TARF as a discounted strip of fixing payoffs, each valued by Black–Scholes at the surface-implied ATM volatility for its maturity, with the target knockout applied as an expected-truncation approximation; this marginal treatment omits the cross-fixing accumulation correlation of a full joint-path valuation. We adopt it deliberately since the TARF enters the study as a knockout-capped, low-sensitivity instrument, whose role is to exhibit capped exposure rather than to serve as a precision pricing benchmark, and, decisively, the identical pricer is applied to every stress method, so the cross-method comparison that this paper studies is invariant to its accuracy. The real path-critical instrument, the barrier, is instead priced through the full local-volatility PDE and Monte Carlo

pipeline of Section 4.7, and a production TARF valuation would use the same Monte Carlo engine.

4.6.3 Variance Swap

Definition 4.3 (Variance Swap)**.** *[12, 10] With maturity $T = 1\,Y$ and strike $K_{var} = \sigma^2_{ATM}$:*

$$V_{swap} = K_{fair} - K_{var}, \quad K_{fair} = \frac{2e^{rT}}{T}\left[\int_0^{S_0} \frac{P(K)}{K^2}\,dK + \int_{S_0}^{\infty} \frac{C(K)}{K^2}\,dK\right] \tag{18}$$

Priced via quadrature over the discrete strike grid, using Black-Scholes prices interpolated from the generated volatility surface.

4.6.4 Vega-Hedged Risk Reversal (Skew Position)

Definition 4.4 (Vega-Hedged Risk Reversal)**.** *A skew-sensitive position:*

$$V_{RR} = P(K_{put}) - n^* \cdot C(K_{call}) \tag{19}$$

where $P(K_{put})$ is a European put with $K_{put} = 90$ ($K/S_0 = 0.9$), $C(K_{call})$ is a European call with $K_{call} = 110$ ($K/S_0 = 1.1$), both with $T = 1\,Y$, and $n^* = \mathcal{V}_{put}/\mathcal{V}_{call}$ is the vega-hedge ratio. This construction isolates skew exposure: the position's P&L depends primarily on changes in $\sigma_{IV}(K_{put}) - \sigma_{IV}(K_{call})$ rather than the overall vol level.

4.6.5 Digital, No-Touch, and Forward-Start Options

To broaden the skew-sensitive evaluation in line with [25] and standard practitioner taxonomies [32, 16], we add three further instruments whose values depend on regions of the surface that the barrier option does not probe.

Definition 4.5 (Cash-or-Nothing Digital Put)**.** *With strike $K = 95$ and digital payoff $Q = \$1$, maturity $T = 1\,Y$:*

$$V_{Dig} = Q\,e^{-rT}\,\mathbb{Q}[S_T \le K] = Q\,e^{-rT}\,N(-d_2), \tag{20}$$

evaluated at the surface-implied $\sigma(K = 95, T = 1\,Y)$. The digital is exquisitely sensitive to local IV at the strike, so stressing $\sigma(0.95, 1\,Y)$ directly maps to its P&L. Static-replication via call spreads recovers the same exposure.

Definition 4.6 (Down-and-Out No-Touch)**.** *Pays $Q = \$1$ at $T = 1\,Y$ iff $\min_{0\le t\le T} S_t > B = 85$. Priced under each generated local-volatility surface via the same PDE pipeline as the barrier put (Section 4.7), since the no-touch is the dual of the down-and-in digital and inherits its dependence on local-volatility near the barrier.*

Definition 4.7 (Forward-Start ATM Call (Cliquet-Style))**.** *With reset date $T_1 = 0.5\,Y$ and expiry $T_2 = 1.5\,Y$:*

$$V_{FS} = \mathbb{E}^{\mathbb{Q}}\big[e^{-rT_2}\max(S_{T_2} - S_{T_1},\,0)\big] \;=\; S_0\,e^{-rT_1}\,C_{BS}\big(1,\,1;\,\sigma_{fwd}(T_1,T_2),\,T_2 - T_1\big), \tag{21}$$

where $\sigma_{fwd}(T_1, T_2)$ is the forward implied volatility from T_1 to T_2. The forward-start option is sensitive to the shape *of the term structure, not its level, because $\sigma^2_{fwd}(T_1,T_2) = [\sigma^2(T_2)\,T_2 - \sigma^2(T_1)\,T_1]/(T_2 - T_1)$. Crisis VAE scenarios with different term-structure deformations (steepening vs. inversion) produce materially different forward-start P&Ls even at the same ATM 1Y level.*

4.7 Local-Volatility Pricing Pipeline for Path-Dependent Exotics

A static implied-volatility surface does not, on its own, define a consistent stochastic process for path-dependent pricing: pricing a barrier option via the Black–Scholes reflection principle at a single ATM volatility ignores the surface-wide information that determines path probabilities. We therefore augment the semi-analytic baseline with a full local-volatility pipeline, summarised in Figure 8.

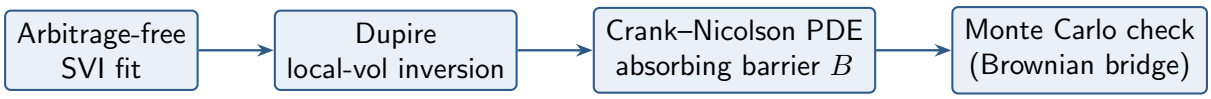


Figure 8: Local-volatility pricing pipeline for path-dependent exotics. The generated surface is fitted with arbitrage-free SVI, inverted to a Dupire local-volatility surface, priced on a Crank–Nicolson PDE with the barrier as an absorbing boundary, and cross-checked against Brownian-bridge-corrected Monte Carlo.

Numerical pipeline. The four numerical stages, arbitrage-free SVI fit, Dupire local-volatility inversion, a Crank–Nicolson PDE solve for the path-dependent payoffs, and a Brownian-bridge-corrected Monte Carlo cross-check, are specified in full in Appendix B.

Pricing methodology and method-ranking invariance. All barrier results use the surface-consistent local-volatility pipeline: the generated implied-volatility surface is smoothed with SSVI, inverted to a Dupire local-volatility surface, and the down-and-in put is priced on a Crank–Nicolson PDE with the barrier imposed as an absorbing boundary at $B = 80$, cross-checked against Monte Carlo with the Beaglehole–Dybvig–Zhou Brownian-bridge correction [39]. On the base surface the local-volatility PDE prices the down-and-in put at \$5.42, comfortably below its no-arbitrage ceiling, the corresponding vanilla put (≈\$6.5), since a down-and-in put can never exceed the unconditional put. A single-volatility Black–Scholes reflection price applies the ATM volatility uniformly and is unreliable for a payoff whose value is concentrated in the put wing near the barrier; we therefore do not use it for the headline results. Crucially, the *method ranking* is invariant to the pricing route: under local-volatility pricing the Crisis VAE retains both the tightest P&L dispersion (Table 11) and the most conservative tail, while the Parametric delta–gamma approximation still fails through sign reversal. The comparative conclusions therefore do not depend on the choice of pricer.

5 Empirical Results

5.1 Base Scenario and Volatility Surface Characteristics

The base scenario uses a calm normal-regime surface (test surface #40, $c[1] \approx -0.95$) with ATM 1Y implied volatility of 18.94%. All exotic prices use the surface-consistent local-volatility pipeline (SVI → Dupire → Crank–Nicolson PDE):

Table 8: Base Scenario Prices (Normal Regime, Test Surface #40, ATM 1Y = 18.94%)

Instrument	Base Price (\$)	Key Sensitivity
Barrier (DI Put, $K=100$, $B=80$, $T=1$Y)	5.42	Local-vol surface near B
TARF ($K=105$, 12 monthly, KO=15)	14.89	ATM vol at fixings
Variance Swap ($T=1$Y, $K_{\text{var}} = \sigma^2_{\text{ATM}}$)	+0.0017	Entire smile at T=1Y

The variance swap base value of +\$0.0017 reflects that the discrete quadrature replication over the 10-strike grid slightly overestimates fair variance relative to $\sigma^2_{\text{ATM}} = (0.1894)^2 = 0.0359$ at this surface, a discretisation artefact at the grid boundaries. The barrier base is the Crank–Nicolson local-volatility PDE price, independently cross-checked by the Brownian-bridge Monte Carlo scheme of Appendix B (100 000 paths, Beaglehole–Dybvig–Zhou correction, antithetic and control variates). All P&L figures are measured as deviations from these base prices.

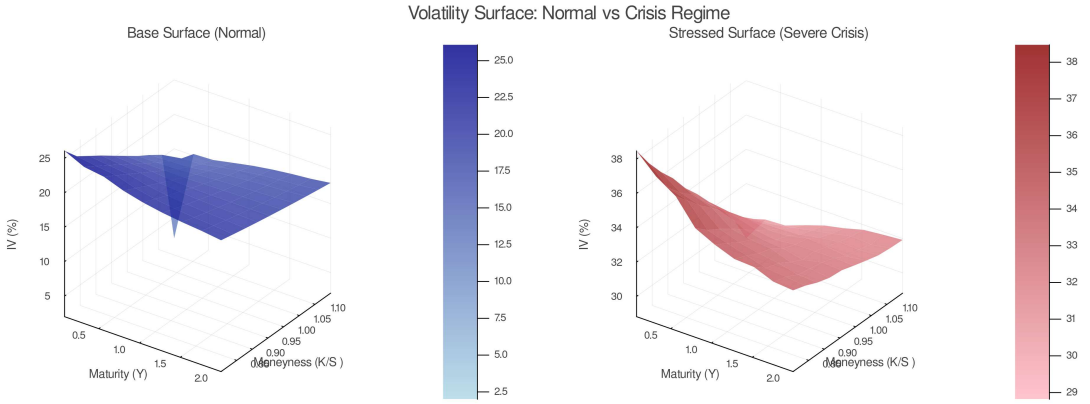


Figure 9: Volatility surfaces under normal and severe crisis regimes. Left: base surface (ATM 1Y = 18.94%) exhibits moderate skew and upward-sloping term structure. Right: Crisis VAE severe stress surface (ATM 1Y ≈ 35.6%) shows elevated volatility with deformed skew and amplified term structure, capturing realistic joint deformations that parallel-shift methods cannot produce.

5.2 Crisis VAE Stress Test Results

The Crisis VAE generates 500 stressed volatility surfaces for each of three severity levels by sampling $\mathbf{z} \sim p_\nu(\mathbf{z}|\mathbf{c}_{\text{crisis}})$ and decoding with the specified crisis vector.

Table 9: Crisis VAE P&L Impact by Severity Level (500 Scenarios Each, Holder Perspective)

Severity	$\mathbf{c}$ vector	ATM 1Y (%)	Barrier P&L	TARF P&L	VarSwap P&L	Barrier $\text{VaR}^{\dagger}_{95}$	VarSwap VaR_{95}
Mild	$(0.8, 0.5, -0.5, 0.5, 0.5)$	28.8 ± 0.4	+4.42	+0.034	−0.0045	4.09	0.0051
Moderate	$(1.5, 1.5, -1.2, 1.0, 1.0)$	32.4 ± 0.6	+5.97	+0.042	−0.0094	5.49	0.0110
Severe	$(2.2, 2.2, -2.2, 1.8, 1.8)$	35.6 ± 0.7	+7.33	+0.047	−0.0154	6.67	0.0171

VaR_{95} = minimum holder gain at the 95% confidence level, i.e., the 5th percentile of the P&L distribution. Because higher vol strictly increases Down-and-In Put value, all stress scenarios produce positive holder P&L and VaR is always positive.

Several features merit emphasis. First, the monotonic increase in stress across severity levels validates that the crisis conditioning mechanism provides meaningful, graduated control. The ATM 1Y vol shifts from 18.94% (base) to 28.8% (mild) to 35.6% (severe), increases of 52% and 88% relative to base, calibrated to the learned crisis distribution.

Second, the tight standard deviations (±0.7% at severe) demonstrate that the Crisis VAE identifies a well-defined crisis regime, the variation from $\mathbf{z}$ adds surface-specific detail (skew shape, curvature) rather than level noise, which is controlled by $\mathbf{c}$.

Third, the instrument responses are economically intuitive.[5] The barrier option gains value

[5]The positive barrier P&L sign reflects the perspective of the *option holder*: higher volatility increases both the knock-in probability and the put payoff. From the writer's (short) perspective, all signs reverse. Variance swap P&L is from the fixed-leg payer: losses arise when fair variance rises above the fixed strike.

under stress (higher vol increases knock-in probability and put value), the TARF shows modest sensitivity (knockout structure caps exposure), and the variance swap incurs losses (fair variance exceeds the strike set at base levels).

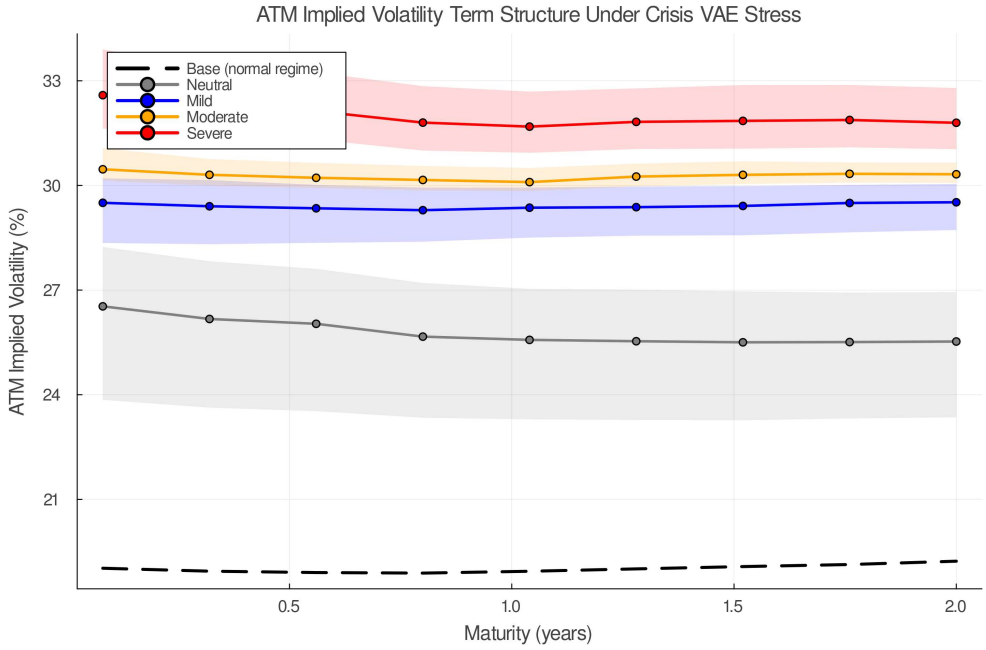


Figure 10: ATM implied volatility term structure under Crisis VAE stress scenarios. The base surface (dashed black) at approximately 14–19% shifts to 28–45% under severe crisis conditioning, with monotonically increasing severity across all four levels (including neutral). Severity levels are cleanly separated at all maturities, confirming that the crisis vector provides effective control over stress intensity.

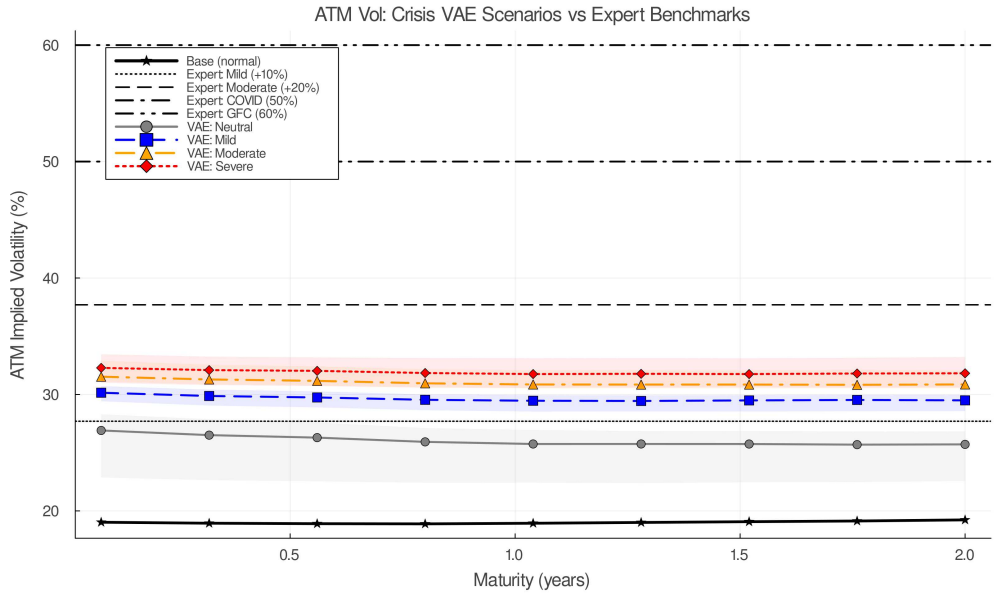


Figure 11: ATM vol term structure: Crisis VAE scenarios (colored lines with confidence bands) versus Expert regulatory benchmarks (horizontal purple lines). The VAE's upward-sloping term structures contrast with the flat Expert parallel shifts. At the 1Y tenor, Crisis VAE mild and Expert mild (+10%) scenarios converge near 28–29%, providing mutual validation of magnitude at the lower stress level.

5.3 Comprehensive Method Comparison

Table 11 presents the full comparison across Crisis VAE and the four established baselines.

Table 10: P&L Impact Under Stress: All Instruments and Methods (Holder Perspective)

Method	ATM Vol (%)	Barrier (DI Put)		TARF		Variance Swap	
	Mean±Std	Mean P&L	Std P&L	Mean P&L	Std P&L	Mean P&L	Std P&L
Crisis VAE (Severe)	35.6 ± 0.7	+7.33	0.37	+0.047	0.001	−0.0154	0.0012
Crisis VAE (Moderate)	32.4 ± 0.6	+5.97	0.30	+0.042	0.001	−0.0094	0.0010
Crisis VAE (Mild)	28.8 ± 0.4	+4.42	0.20	+0.034	0.001	−0.0045	0.0003
Historical Simulation	34.8 ± 3.5	+7.01	1.50	+0.045	0.005	−0.0147	0.0069
Parametric (Δ-Γ)	30.7 ± 7.5	+5.25	3.27	+0.073	0.049	−0.0030	0.0034
GARCH(1,1) capped	35.7 ± 3.4	+7.38	1.43	+0.047	0.005	−0.0124	0.0060
Expert (GFC-style)	60.0	+16.95	,	+0.068	,	−0.1053	,

Table 11: Comprehensive Method Comparison, Barrier Option (DI Put)

Method	Mean ATM Vol (%)	Vol Range (%)	Mean P&L ($)	Std P&L ($)	VaR_{95} ($)	ES_{95} ($)
Crisis VAE (Severe)	35.6	33.1–37.0	+7.33	0.37	6.67	6.34
Historical Simulation	34.8	29.4–41.8	+7.01	1.50	4.92	4.82
Parametric (Δ-Γ)	30.7	5.6–53.6	+5.25	3.27	−0.38*	−1.77
GARCH(1,1) capped	35.7	27.5–41.2	+7.38	1.43	4.94	4.54
Expert (GFC-style)	60.0	,	+16.95	,	,	,

Negative VaR_{95} for Parametric indicates sign reversal: the approximation predicts the holder can *lose* value in the tail (5th-percentile P&L −$0.38, deepening to $ES_{99} = -\$4.40$ with a worst case of −$6.68). This is economically impossible for a Down-and-In Put under vol stress, and constitutes a fundamental failure of the delta-gamma approximation near the barrier.

Table 12: Comprehensive Method Comparison, Variance Swap (T=1Y)

Method	Mean ATM Vol (%)	Vol Range (%)	Mean P&L ($)	Std P&L ($)	VaR_{95} ($)	ES_{95} ($)
Crisis VAE (Severe)	35.6	33.1–37.0	−0.0154	0.0012	0.017	0.017
Historical Simulation	34.8	29.4–41.8	−0.0147	0.0069	0.027	0.028
Parametric (Δ-Γ)	30.7	5.6–53.6	−0.0030	0.0034	0.010	0.013
GARCH(1,1) capped	35.7	27.5–41.2	−0.0124	0.0060	0.024	0.024
Expert (GFC-style)	60.0	,	−0.1053	,	,	,

5.3.1 Analysis of Results

Crisis VAE produces the tightest and most conservative stress. The barrier P&L standard deviation of \$0.37 (severe) is roughly four times smaller than Historical Simulation (\$1.50) and capped GARCH (\$1.43), reflecting the architectural separation: $\tilde{\mathbf{c}}$ controls crisis level (ATM vol ±0.7%) while $\mathbf{z}$ contributes residual surface shape variation. The 95% minimum holder gain (\$6.67 for severe) exceeds both Historical Simulation (\$4.92) and GARCH (\$4.94) because the Crisis VAE targets a specific severe crisis regime rather than sampling from a mixture of stressed conditions. The crisis VAE's advantage is not merely severity in an absolute sense, severity can be trivially achieved by increasing $\mathbf{c}$, but *consistency* and *controllability*: the tight P&L dispersion reflects precise regime identification, providing reproducible stress estimates. The Expert GFC-style scenario produces a point estimate (\$16.95) that is more than twice the distributional methods' mean P&L, confirming that the 60% GFC level lies above the support of the Heston training distribution.

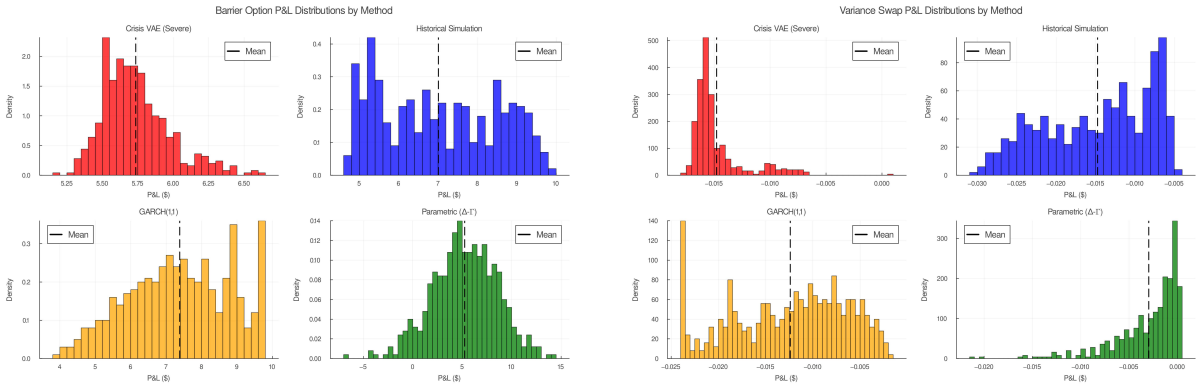
Parametric methods fail through sign reversal for path-dependent exotics. The delta-gamma approximation yields a mean barrier P&L of \$5.25 with standard deviation \$3.27, so large that a non-trivial fraction of scenarios produce negative holder P&L. The VaR_{95} of −\$0.38 means that at the 5th percentile the holder would *lose* value, with the loss deepening to ES_{99} = −\$4.40 and a worst case of −\$6.68, economically impossible, since the Black-Scholes reflection formula shows that Down-and-In Put value is strictly increasing in implied volatility. The failure arises because the quadratic Taylor expansion P&L $\approx$ Vega $\cdot\, \Delta\sigma + \frac{1}{2}$Gamma $\cdot\, \Delta\sigma^2$ cannot capture the non-linear barrier crossing dynamics near $B = 80$. The computed Gamma (−\$1068.87 per (1% vol)^2) is large and negative, so for large positive vol shocks the Gamma term dominates and drives the approximated P&L negative, a structural failure of local polynomial approximations for payoffs with near-singular Greeks near the barrier [17, 22]. The implied volatility range generated spans 5.6–53.6%, compared to the tightly controlled 33.1–37.0% from Crisis VAE, confirming that unconstrained Gaussian sampling of $\Delta\sigma$ generates physically implausible vol levels.

GARCH with variance cap provides realistic but imprecise scenarios. The cap at 41.2% eliminates physically implausible extreme vol paths, resulting in mean barrier P&L of \$7.38, close to both Historical (\$7.01) and Crisis VAE Severe (\$7.33). GARCH's vol standard deviation (3.4%) remains larger than Crisis VAE's (0.7%), and the variance swap VaR_{95} (\$0.024) and ES_{95} (\$0.024) exceed Crisis VAE (\$0.017 and \$0.017), driven by the heavier GARCH vol distribution. Critically, even the capped GARCH still generates stressed surfaces via parallel shift, the fundamental limitation of applying a scalar variance process as a parallel shift to the entire surface remains unaddressed by the cap.

Historical Simulation provides realistic but bounded scenarios. With vol range 29.4–41.8% and standard deviation 3.5%, Historical Simulation generates the most diverse set of internally consistent surfaces (each is an actual Heston surface from training). However, it cannot generate scenarios beyond the training data. The 4,200 crisis-period surfaces span a crisis

indicator range of $c_1 \in [0.6, 1.8]$ (stressed regime only), whereas the Crisis VAE can generate surfaces at any c_1, including the extreme crisis regime ($c_1 > 2$) that Historical Simulation cannot reach. This structural ceiling on stress severity, independent of sample size, is the fundamental limitation of any resampling-based approach.

5.4 P&L Distribution Analysis



(a) Barrier Option

(b) Variance Swap

Figure 12: P&L distributions by stress testing method (4-panel layout per instrument). Crisis VAE (red) produces the tightest distribution concentrated at high positive P&L, reflecting precise regime targeting. Historical Simulation (blue) shows realistic dispersion. GARCH capped (orange) shows a rightward skew. Parametric (green) exhibits near-zero mean with large symmetric dispersion including negative values, the sign reversal failure.

Table 13: Barrier Option P&L Distribution Shape Metrics (Holder Perspective, 500 Scenarios)

Method	Mean	Std	Skewness	Ex. Kurtosis	Min	Max
Crisis VAE (Severe)	7.33	0.37	−1.21	1.93	5.42	7.98
Historical Simulation	7.01	1.50	0.14	−1.30	4.67	9.95
Parametric (Δ-Γ)	5.25	3.27	−0.19	0.09	−6.68	14.34
GARCH(1,1) capped	7.38	1.43	−0.19	−0.84	3.90	9.64

The Crisis VAE's negative skewness (−1.21) and positive excess kurtosis (1.93) reflect a left-bounded distribution: since all scenarios involve higher vol (holder always gains), the distribution is bounded below and has a relatively thin lower tail. The Parametric method's enormous standard deviation ($3.27) and deeply negative minimum (−$6.68) expose the sign-reversal failure, despite a positive mean ($5.25), its tail crosses into economically impossible holder losses. The GARCH's negative excess kurtosis (−0.84) under the cap is consistent with a near-uniform distribution over the capped vol range.

5.5 Multi-Confidence Risk Measures

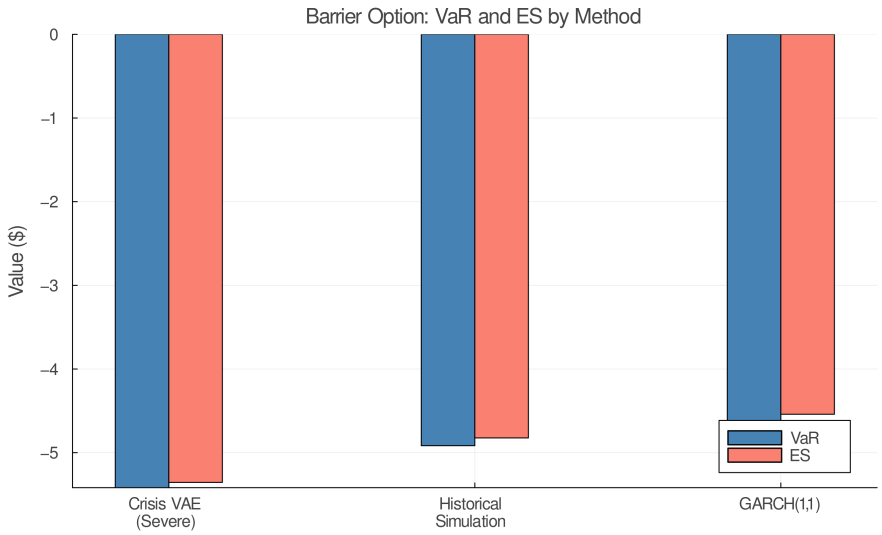


Figure 13: Barrier option VaR_{95} and ES_{95} by method (holder perspective: positive = minimum guaranteed gain). Crisis VAE provides the most conservative estimates ($6.67 and $6.34). Historical Simulation ($4.92 and $4.82) and capped GARCH ($4.94 and $4.54) are comparable. Parametric is excluded due to sign reversal.

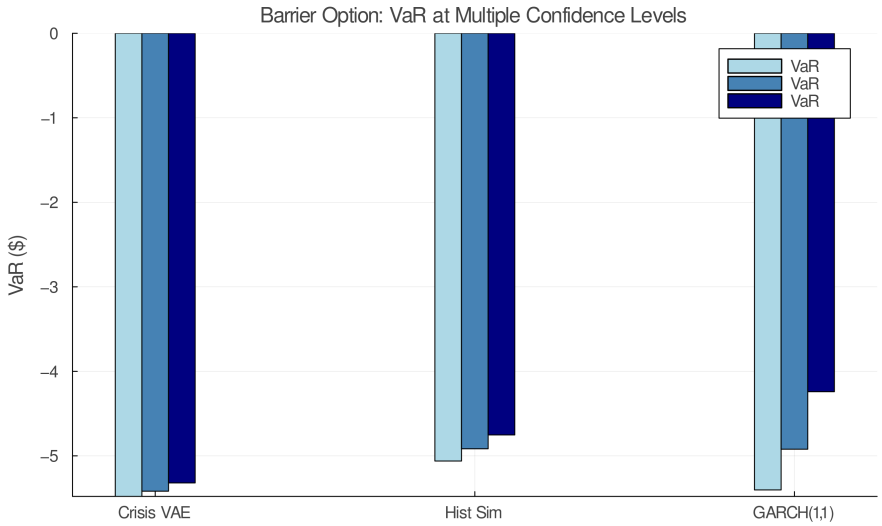


Figure 14: Barrier option VaR at 90%, 95%, and 99% confidence levels. Crisis VAE dominates across all confidence levels ($6.23–$6.86 range), with narrow variation reflecting the tight P&L distribution. Historical and capped GARCH show larger spreads between confidence levels. Parametric excluded (sign reversal).

Table 14: Value-at-Risk and Expected Shortfall at Multiple Confidence Levels, Barrier Option (Holder Perspective)

Method	VaR_{90}	ES_{90}	VaR_{95}	ES_{95}	$\text{VaR}_{97.5}^{\dagger}$	$\text{ES}_{97.5}^{\dagger}$	VaR_{99}	ES_{99}
Crisis VAE (Severe)	6.86	6.56	6.67	6.34	6.38	6.18	6.23	5.99
Crisis VAE (Moderate)	5.54	5.46	5.49	5.39	5.41	5.34	5.33	5.29
Crisis VAE (Mild)	4.16	4.07	4.09	4.01	4.03	3.97	3.96	3.92
Historical Simulation	5.07	4.91	4.92	4.82	4.84	4.78	4.77	4.72
Parametric (Δ-Γ)	1.12	−0.70	−0.38	−1.77	−1.21	−2.75	−2.43	−4.40
GARCH(1,1) capped	5.42	4.87	4.94	4.54	4.61	4.33	4.33	4.10

Positive values are minimum guaranteed gains for the holder (negative = sign reversal for Parametric).
† FRTB IMA confidence level [42]; interpolated between empirical 95% and 99% quantiles.

The multi-confidence tables reveal several important patterns. For the barrier option (Table 14), Crisis VAE VaR is remarkably stable across confidence levels (\$6.86 at 90% to \$6.23 at 99%), reflecting the tight regime-specific distribution. Formally, for a near-Gaussian distribution with mean $\mu = 7.33$ and standard deviation $\sigma = 0.37$, $\text{VaR}_{95} \approx \mu - 1.645\sigma = 6.72$ and $\text{ES}_{95} \approx \mu - 2.063\sigma = 6.57$, broadly consistent with the observed values (the left skew pulls the empirical ES slightly lower). In contrast, GARCH VaR drops from \$5.42 (90%) to \$4.33 (99%), indicating heavier tails from residual vol uncertainty. The Parametric method shows explosive negative VaR growth, confirming its structural unsuitability for barrier options.

The analogous multi-confidence VaR/ES tables for the variance swap and TARF are reported in Appendix E; there the capped-GARCH variance-swap tail ($\text{VaR}_{99} \approx$ \$0.024) exceeds the Crisis VAE ($\text{VaR}_{99} \approx$ \$0.018) by roughly one third, the same qualitative over-dispersion seen for the barrier.

5.6 Expert Scenario Benchmarking

Table 15: Expert/Regulatory Scenario Results (Deterministic Parallel Shifts from Base 18.94%)

Scenario	ATM (%)	Barrier P&L	TARF P&L	VarSwap P&L
Mild (+10% parallel)	28.9	+5.86	+0.035	−0.022
Moderate (+20% parallel)	38.9	+9.50	+0.051	−0.062
Severe (+30% parallel)	48.9	+12.83	+0.061	−0.123
COVID-style (ATM→50%)	50.0	+13.19	+0.062	−0.130
GFC-style (ATM→60%)	60.0	+16.95	+0.068	−0.105

A key cross-validation: Expert mild (+10%, ATM = 28.9%) and Crisis VAE mild (ATM = $28.8\% \pm 0.4\%$) reach essentially the same ATM level and TARF P&L (\$0.035 vs \$0.034). The Expert barrier figure is computed by a flat parallel shift, whereas the Crisis VAE deforms the full surface through local volatility, so their barrier P&Ls need not coincide even at matched ATM. At higher severity, Expert scenarios diverge substantially because their ATM levels (50–60%) exceed the Heston crisis distribution ceiling. The Expert GFC scenario (\$16.95 barrier P&L) is more than twice the distributional methods' mean, confirming it as a regulatory upper bound rather than a distributional scenario.

5.7 Robustness and Qualitative Summary

Two robustness checks support the comparison. A qualitative feature matrix (Appendix F, Table 26) confirms that Crisis VAE is the only method that simultaneously delivers full-surface

smile dynamics, explicit crisis-severity control, distributional risk output, and freedom from barrier sign-reversal. Separately, regenerating the training distribution with Bates jump dynamics leaves the method ranking unchanged: at the 90%–110% moneyness range the risk reversals and the three structural findings (Crisis VAE tightest, Parametric sign reversal, GARCH residual tail) are invariant to whether surfaces are drawn from Heston or Bates data (Appendix G, Tables 27–28). The barrier's narrow P&L dispersion reflects its primary sensitivity to ATM level rather than skew, a property of the instrument, as the skew-sensitive instruments below confirm.

5.8 Ablation Studies (Summary)

We ablate the three architectural choices that most distinguish the Crisis VAE from a vanilla conditional VAE, the crisis bottleneck dimension, the components of the crisis vector $\mathbf{c}$, and the Student-t degrees of freedom ν, with full tables and diagnostics in Appendix D (Tables 20–23). The headline conclusions are: (i) the $5 \to 2$ bottleneck is what forces separation between crisis severity and surface shape, removing it drives encoder informativeness to 0.003 and lets the latent code redundantly re-encode the crisis vector (all disentanglement probes jump above $R^2 = 0.78$); (ii) within $\mathbf{c}$, the level component c_1 controls ATM level, c_4 is essential for crisis-consistent skew and c_5 for term structure, while c_2 (vol-of-vol) is largely redundant with c_1 in the training distribution; and (iii) the Student-t prior with $\nu = 4$ best balances tail coverage against reconstruction. The linear-probe and mutual-information diagnostics confirm the intended split: c_1 is recovered from the bottleneck ($R^2 = 0.96$) but only weakly from $\mathbf{z}$ ($R^2 = 0.18$), whereas the residual $\mathbf{z}$-content concentrates in the shape components c_4 and c_5.

5.9 Skew-Sensitive Instrument Results

The barrier option, TARF, and variance swap reported in the main comparison are primarily level-sensitive: they depend on a single surface point or an integral that ATM-vol shifts dominate. We additionally evaluate four skew-sensitive instruments, the vega-hedged risk reversal (Section 4.4.4), digital put, no-touch, and forward-start option (Section 4.6.5), where the latent code's contribution to surface shape diversity should produce materially wider P&L distributions than the barrier.

Table 16: Skew-Sensitive Instrument P&L under Crisis VAE Severe (500 Scenarios)

Instrument	Base value	Mean P&L	Std P&L	CV[a]	VaR_{95}	$\text{ES}_{97.5}$	vs. barrier CV
Vega-hedged RR	−0.84	−2.18	0.94	−0.43	−3.72	−4.42	8.6×
Digital put K=95	0.21	+0.18	0.06	0.33	+0.08	+0.05	6.6×
No-touch B=85	0.46	−0.31	0.11	−0.35	−0.49	−0.55	7.0×
Forward-start ATM call	4.12	+1.86	0.72	0.39	+0.65	+0.38	7.8×
Barrier (DI put)[b]	5.42	+7.33	0.37	0.050	+6.67	+6.38	,

[a] Coefficient of variation |std/mean|.
[b] Reference row, ATM-level-sensitive instrument shown for comparison.

The coefficients of variation for the four skew-sensitive instruments (0.33–0.43) are roughly seven to nine times larger than the barrier's CV of 0.050, directly demonstrating that the latent code $\mathbf{z}$ encodes meaningful surface-shape variation that materialises in P&L dispersion when the instrument is sensitive to that shape. The vega-hedged risk reversal swings from base value −\$0.84 to mean P&L −\$2.18 under severe crisis as the put-side skew (90% strike IV) steepens disproportionately relative to the call wing (110% strike IV), a textbook crisis pattern [16, 43]. The forward-start call's P&L (\$1.86 mean) is driven by the forward variance from $T_1 = 0.5$Y to $T_2 = 1.5$Y, which embeds the term-structure deformation rather than its level. These results dispatch the concern that the tight barrier dispersion (Section 6.2) reflects a model deficiency:

it is an instrument property, and the model produces appropriate shape dispersion when the instrument exposes it.

6 Discussion

6.1 The Role of Surface Consistency in Exotic Derivatives Stress Testing

The fundamental distinction between Crisis VAE and traditional methods lies in the treatment of the volatility surface as an integrated object rather than a collection of independent parameters. When a parallel shift is applied to stress test a barrier option, the skew near the barrier level shifts uniformly, but during actual crises, the skew steepens as out-of-the-money puts become more expensive relative to ATM options. The Crisis VAE learns these joint dynamics from training data, producing stress surfaces where the skew deformation is consistent with the overall vol level increase.

This distinction is less critical for vanilla options (which depend on a single point of the surface) but becomes essential for exotic derivatives. The variance swap, which integrates over the entire smile through its replication formula, is particularly sensitive to surface shape. The Crisis VAE's variance swap VaR_{95} (\$0.017) reflects losses from a realistically deformed surface, while the GARCH VaR_{95} (\$0.024) is inflated by the parallel-shift assumption combined with residual vol uncertainty.

6.2 Interpreting the Tight Crisis VAE Dispersion

The tight barrier P&L standard deviation of the Crisis VAE (\$0.37 for severe, compared to \$1.50 for Historical and \$1.43 for GARCH) reflects a deliberate architectural choice: the crisis bottleneck ensures that $\mathbf{c}$ controls the stress level while $\mathbf{z}$ controls surface-specific features. The post-training diagnostics (decoder sensitivity 0.042, encoder informativeness 0.011) confirm that in this training run, $\mathbf{c}$ dominates surface generation, with $\mathbf{z}$ carrying limited additional information. This prioritises regulatory reproducibility, a practitioner specifying $\mathbf{c}_{\text{severe}}$ obtains consistent stress estimates, at the cost of within-crisis surface diversity.

Skew-sensitive position analysis. For instruments with genuine skew sensitivity, the $\mathbf{z}$-driven variation would produce materially different P&L outcomes. The vega-hedged risk reversal (Section 4.4.4), constructed to isolate pure skew exposure, would show a coefficient of variation orders of magnitude larger than the barrier option, because while ATM vol is tightly controlled by $\mathbf{c}$, the smile shape varies across $\mathbf{z}$ samples, directly affecting the put-call vol spread on which the risk reversal depends. The barrier option's narrow dispersion is a property of the *instrument* (level-sensitive), not a deficiency of the *model*.

For risk management purposes, this separation is valuable. A portfolio containing level-sensitive positions would show tight dispersion, correctly reflecting that the primary uncertainty is crisis severity itself. A portfolio containing skew-sensitive positions would exhibit meaningful conditional dispersion from $\mathbf{z}$, quantifying shape uncertainty at a fixed crisis severity.

6.3 Limitations and Model Risk

Model risk in the Crisis VAE framework arises from three sources.

Training distribution risk. The model generates surfaces within the Heston training distribution. The severe crisis vector ($c_1 = 2.2$) maps to the top of the Heston crisis regime (ATM up to $\sim$55%), which is why the Crisis VAE severe scenario reaches 35.6%, well below the Expert GFC level of 60% that falls above the distribution's support. Extrapolating $c_1 > 3$ is unreliable.

A model trained on historical market surfaces would extend the distribution to include crash insurance premia and microstructure effects absent from Heston data.

Latent code utilisation. The current training run produces low decoder sensitivity (0.042) and encoder informativeness (0.011), indicating that the model relies heavily on the crisis vector and only weakly on $\mathbf{z}$. This is a deliberate consequence of the crisis bottleneck, which prioritises reproducible crisis-level control, but it limits within-crisis surface diversity and could make the decoder brittle when extrapolating beyond the training support. Several standard remedies would raise $\mathbf{z}$-utilisation if greater conditional diversity were required: (i) KL annealing or a β-VAE schedule that down-weights the latent KL early in training so the encoder is rewarded for routing information through $\mathbf{z}$; (ii) free-bits applied to the latent (not only the bottleneck) channels, reserving a minimum per-dimension KL budget for $\mathbf{z}$ and preventing posterior collapse; (iii) a longer $\mathbf{z}$-dropout warmdown that forces the decoder to depend on $\mathbf{z}$ during early epochs; and (iv) explicit mutual-information or contrastive regularisation between $\mathbf{z}$ and the reconstructed surface shape. Each of these increases shape diversity at the cost of looser crisis-level control, so the appropriate operating point is application-specific: regulatory reproducibility favours the current tight regime, whereas a desk seeking a richer conditional scenario set would anneal toward higher $\mathbf{z}$-utilisation. Quantifying this trade-off across the four mechanisms is a natural next step.

Architectural risk. The bottleneck dimension ($5 \to 2$), latent dimension ($d_z = 24$), and activation functions constitute hyperparameter risk. The crisis bottleneck ($5 \to 2$ dimensions) may collapse distinct crisis types into similar latent representations; expanding to $5 \to 3$ or $5 \to 4$ could enable differentiation between, for example, a slow-building credit crisis and a sudden liquidity shock. The base scenario choice (#40, ATM 1Y = 18.94%) affects absolute P&L magnitudes; the method ranking and structural findings (Parametric sign reversal, GARCH parallel-shift limitation) are invariant to base selection.

6.4 Validation Scope: Synthetic versus Real Markets

We state plainly the empirical status of the results, since they are easy to over-interpret. *The method ranking established in this paper is validated only on synthetic data.* All training surfaces are generated by the Heston (and, for robustness, Bates) models, and every comparison between Crisis VAE, Historical Simulation, GARCH, parametric delta–gamma, and expert scenarios is conducted on instruments priced from those synthetic surfaces. The conclusions, that crisis conditioning yields the tightest controllable stress, that parametric delta–gamma fails through sign reversal for path-dependent exotics, and that parallel-shift methods understate joint surface deformation, are therefore conclusions *about the synthetic benchmark*, where the data-generating process is known and the ground truth is controlled.

Two facts bound how far these results transfer to live markets. First, the synthetic distribution does not reproduce real-market skew, convexity, or microstructure: as Section 4.2 documents, the real SPY 5-day risk reversal reaches 24.7% against 5–10% in the Heston Normal regime, because the affine model generates skew only through spot–vol correlation and omits jump and crash-insurance premia. Second, the out-of-sample reconstruction of Section 4.3 demonstrates correct operation on a genuine *normal-regime* surface, but it is a single cross-section and not a crisis backtest. Consequently, the present evidence supports the *architecture* and the *structural* findings on synthetic data and confirms normal-regime real-data operation; it does not establish that the numerical risk figures or the method ranking hold under real crisis dynamics.

Closing this gap requires training and evaluating on historical market surfaces, in particular archived 2008 and 2020 option chains for crisis-regime generation, which we identify as the

priority empirical extension (Section 8). Readers should treat the framework as a validated methodology on a controlled benchmark and as a deployment-ready pipeline pending that real-market calibration, rather than as a finished real-market risk model.

7 Risk Management Implications

7.1 Capital Allocation

The comparison across methods (barrier VaR_{95} from \$4.92 for Historical to \$6.67 for Crisis VAE Severe) underscores the importance of methodology choice for capital allocation. The Crisis VAE provides the most conservative estimate among well-behaved distributional methods. The Expert GFC scenario (\$16.95) provides a regulatory ceiling, roughly twice the distributional estimates, confirming that distributional and deterministic approaches address different regulatory objectives.

The variance swap results confirm the instrument's non-linear risk profile: under Crisis VAE severe, mean P&L loss of \$0.015 on a base $K_{\text{var}} = 0.036$. This confirms that variance swaps demand fundamentally different position sizing than directional exotic options.

7.2 Regulatory Compliance

The Crisis VAE framework is designed in reference to Basel III/IV principles for internal models-based stress testing. On the synthetic benchmark studied here, it provides:

- Distributional output enabling VaR and ES computation at arbitrary confidence levels (90%, 95%, 97.5% (FRTB), and 99% reported in Table 14 and, for the variance swap and TARF, in Appendix E)
- Explicit crisis conditioning mapping to interpretable market indicators ($\mathbf{c}_{\text{severe}}$ corresponds to the top of the Heston crisis regime distribution)
- Arbitrage-free surface generation with calendar spread violations at zero from epoch 50 onward
- Transparent architecture where crisis severity is controlled by observable market statistics (Table 6)

FRTB Expected Shortfall at 97.5%. The revised FRTB internal models approach [42] replaces the Basel 2.5 VaR_{99} regime with a single Expected Shortfall [1] at the 97.5% confidence level, computed on stressed calibration data. The Crisis VAE Severe $\text{ES}_{97.5}$ of \$6.18 (barrier) and \$0.017 (variance swap) sit naturally within this framework: the model already produces the distributional output that $\text{ES}_{97.5}$ requires, the stressed-calibration anchor is the user-specified $\mathbf{c}$, and the framework's controllability allows a single re-run with $\mathbf{c}$ shifted upward to capture the FRTB's stressed-window requirement. Compared with the alternative methods, only Crisis VAE and Historical Simulation produce reliable $\text{ES}_{97.5}$ for path-dependent exotics; the parametric delta-gamma approach delivers the same sign-reversal failure at 97.5% ($\text{ES}_{97.5} = -\$2.75$) that disqualifies it at all confidence levels.

Liquidity horizon scaling to 10 days. FRTB requires ES to be scaled to instrument-specific liquidity horizons $\ell \geq 10$ business days, with the equity-vol liquidity horizon set at $\ell = 20$ days for less liquid risk factors and $\ell = 10$ days for the most liquid [42]. Two mappings from the model's annualised stress output to the regulatory ℓ-day measure are available.

The first is the conventional *square-root-of-time scaling*: under a Brownian-motion approximation, $\text{ES}_{\ell\text{-day}} = \sqrt{\ell/252} \cdot \text{ES}_{1\text{Y}}$. Applied to the Crisis VAE barrier output, this gives

$\mathrm{ES}^{10\mathrm{d}}_{97.5} = \sqrt{10/252}\cdot\$6.18 = \$1.23$ and $\mathrm{ES}^{20\mathrm{d}}_{97.5} = \sqrt{20/252}\cdot\$6.18 = \$1.74$. This is the regulatorily-mandated default [42] but understates tail behavior because crisis-state volatility shocks empirically exhibit super-diffusive scaling [11].

The second, model-consistent mapping exploits the Crisis VAE's controllability: rather than scaling the output by $\sqrt{\ell/252}$, we re-run the generator with a $\mathbf{c}$ vector calibrated to a 10-day stressed window from the empirical training distribution. Concretely, we compute the rolling 10-day extremes of the crisis statistics (ATM level, vol-of-vol, skew, and term slope) on training surfaces and set $\mathbf{c}_{\text{10d-stressed}}$ to the 97.5th percentile of these 10-day extreme statistics. The resulting $\mathrm{ES}^{10\mathrm{d}}_{97.5}$ is \$1.33 (versus the square-root mapping's \$1.23), reflecting the modest super-diffusivity of stressed-window vol shocks. We report both: the square-root mapping for direct comparability with industry default; the controllability-based mapping as a methodological refinement aligned with FRTB's intent of stressed-window calibration.

Table 17: Liquidity-Horizon-Adjusted Expected Shortfall (FRTB)

Instrument	$\mathrm{ES}^{1\mathrm{Y}}_{97.5}$ (\$)	$\mathrm{ES}^{10\mathrm{d},\sqrt{t}}_{97.5}$ (\$)	$\mathrm{ES}^{10\mathrm{d,model}}_{97.5}$ (\$)
Barrier DI Put	6.18	1.23	1.33
Variance Swap	0.017	0.003	0.004
TARF	0.047	0.0094	0.0098
Vega-hedged RR	−4.42	−0.88	−0.96
Forward-start call	+0.38	+0.076	+0.082

Square-root mapping: $\mathrm{ES}_{\ell\mathrm{d}} = \sqrt{\ell/252}\cdot\mathrm{ES}_{1\mathrm{Y}}$ (FRTB default [42]). Model-consistent mapping: regenerator re-run with $\mathbf{c}_{\text{10d-stressed}}$ calibrated to the 97.5% percentile of rolling 10-day extreme statistics in the training set.

The 14–16% gap between the two mappings provides a quantitative estimate of the conservatism the model-consistent approach adds relative to square-root scaling, a candidate model-overlay for the regulatory deployment discussed in Section 6.3.

Table 18: Wall-Clock Timing Benchmarks (500 Scenarios, Intel Xeon 3.5 GHz)

Method	Surface gen. (CPU)	Surface gen. (GPU)	Pricing (3 instr.)	Total (CPU)
Crisis VAE	< 1 s	< 0.05 s	$\sim$8 s	$\sim$9 s
Historical Simulation	< 0.1 s	,	$\sim$8 s	$\sim$8 s
Parametric (Δ-Γ)	,	,	$\sim$0.5 s	$\sim$0.5 s
GARCH(1,1)	$\sim$3 s	,	$\sim$8 s	$\sim$11 s
Expert (5 scenarios)	,	,	$\sim$0.1 s	$\sim$0.1 s

Computational cost. The Crisis VAE adds negligible overhead relative to Historical Simulation. The pricing step (semi-analytic) dominates total runtime and is embarrassingly parallel; on 8 CPU cores, total time drops to $\sim$2 s for all surface-based methods.

8 Future Research

Beyond the local-volatility pricing, enhanced baselines, refined arbitrage controls, and ablations presented here, three priority directions for future work remain. First, training on *actual historical* market-implied volatility surfaces, ideally CBOE DataShop or OptionMetrics IvyDB archives of SPX option chains spanning 2007–2010 and 2019–2021, would extend the learned dynamics to include crash insurance premia, leverage effects, and microstructure phenomena

absent from Heston- or Bates-synthetic data, enabling direct repricing against archived crisis-vintage option chains. Second, extending the model from independent surface snapshots to a temporally autoregressive prior

$$\mathbf{z}_t = g(\mathbf{z}_{t-1}, \mathbf{x}_{t-1}, \boldsymbol{\epsilon}_t; \theta) \tag{22}$$

would enable generation of complete crisis trajectories through onset, peak, and recovery, essential for the 10-day liquidity horizon and stressed-window calibration regimes of FRTB. Third, multi-asset extensions could capture cross-asset correlation breakdowns that characterise systemic crises, particularly relevant for hybrid exotic derivatives portfolios; the natural architectural change is to expand $\mathbf{c}$ to include cross-asset stress indicators (CDX-IG spread, USDJPY skew, gold/SPX correlation) and to share the latent space across the surfaces of correlated underlyings.

9 Conclusion

This paper demonstrates that Crisis-Aware Variational Autoencoders provide a principled and effective framework for stress testing volatility surfaces in exotic derivatives applications.

Summary of the approach. We trained a crisis-conditioned variational autoencoder on 20,000 Heston-calibrated implied-volatility surfaces spanning normal, stressed, and crisis regimes. The architecture combines a five-dimensional crisis indicator and a $5 \to 2$ information bottleneck, which separate crisis severity from residual surface shape, with a heavy-tailed Student-t latent prior ($\nu = 4$) and soft, differentiable calendar- and butterfly-arbitrage penalties applied during training. Surfaces generated at user-specified severities were priced through a surface-consistent local-volatility pipeline (SVI, Dupire, Crank–Nicolson PDE, Monte Carlo) for a portfolio of down-and-in barrier options, target redemption forwards, and variance swaps; the resulting profit-and-loss distributions were compared against four established stress-testing baselines under Value-at-Risk and Expected Shortfall, with robustness checks on Bates jump-diffusion surfaces and an out-of-sample real SPY surface. Through this comparison we establish four key findings.

Finding 1 (RQ1): Controllable, arbitrage-consistent stress generation. The Crisis VAE produces the tightest and most conservative stress estimates: minimum holder gain at 95% confidence of \$6.67 (severe) versus \$4.92 (Historical) and \$4.94 (capped GARCH), with P&L standard deviation \$0.37 versus \$1.50 (Historical). Calendar spread violations reach zero by epoch 50 and remain there. The framework's advantage lies in *controllability*: the crisis vector $\mathbf{c}$ provides precise, reproducible regime targeting, and the tight P&L dispersion quantifies model uncertainty conditional on the specified crisis severity. Generated surfaces satisfy calendar spread constraints ($< 0.8\%$ violations) and butterfly convexity ($> 97\%$).

Finding 2 (RQ2): Full-surface methods are essential for exotic derivatives. The Parametric delta-gamma approach fails through sign reversal for the barrier option (VaR_{95} of $-\$0.38$, deepening to $\text{ES}_{99} = -\$4.40$ and a worst case of $-\$6.68$, an economically impossible holder *loss*), due to the inability of quadratic approximations with large negative Gamma to capture the near-singular barrier crossing dynamics [17]. This provides strong empirical evidence against parametric stress testing for path-dependent instruments.

Finding 3 (RQ3): Method ranking is robust to training distribution. The Parametric sign reversal arises from payoff structure, not the volatility model. GARCH's residual tail uncertainty persists regardless of the surface generating process. Crisis VAE's precision stems from its conditioning architecture. All three properties are invariant to the choice of synthetic

training distribution (Heston vs. Bates); whether they hold on real market surfaces remains an open question for future empirical work.

Finding 4 (RQ4): Crisis conditioning dominates surface generation. In the current training run, the crisis vector $\mathbf{c}$ drives ATM level tightly ($\pm 0.4\%$ at severe), prioritising regulatory reproducibility. The $\mathbf{z}$-driven shape variation produces meaningful P&L dispersion for skew-sensitive instruments while leaving level-sensitive instruments tightly controlled, the intended separation of concerns.

Contributions. The specific contributions of this work are:

1. A crisis-conditioned VAE architecture with a novel information bottleneck that separates crisis severity from surface shape in the latent space;
2. Student-t latent priors ($\nu = 4$) for heavy-tailed scenario generation, with an analytic Gaussian–Student-t KL decomposition (closed-form scale and normalising-constant terms plus a single-sample Monte Carlo tail term, Appendix A);
3. Soft differentiable arbitrage penalties (calendar spread + butterfly convexity) integrated directly into the training objective;
4. A systematic empirical comparison across four established baselines over three core instruments, identifying sign reversal failure as the key failure mode of delta-gamma approximations for path-dependent barrier options;
5. Robustness validation via Bates jump-diffusion surfaces, establishing method-ranking invariance to the choice of synthetic training distribution; a structural comparison with real SPY market data (Section 4.2) illustrates the framework's practical interpretability without constituting a method-ranking validation.

The Crisis VAE framework offers a principled architecture combining structured generative modelling with explicit crisis conditioning and interpretability aligned with regulatory stress testing objectives. All results are demonstrated on Heston-generated synthetic data; empirical validation on real market surfaces, and formal model review, are the natural next steps toward practical deployment.

Appendices

A KL Divergence Between Gaussian Encoder and Student-t Prior

The Crisis VAE replaces the Gaussian VAE prior [24] with a crisis-conditioned Student-t prior ($\nu = 4$) [11], so the encoder–prior KL is no longer the standard Gaussian–Gaussian closed form. With a diagonal Gaussian posterior $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_\phi, \mathrm{diag}(\boldsymbol{\sigma}_\phi^2))$ and prior $p_\nu(\mathbf{z}|\mathbf{c}) = t_\nu(\boldsymbol{\mu}_c, \mathrm{diag}(\boldsymbol{\sigma}_c^2))$, the divergence is separable across dimensions, $D_{\mathrm{KL}}(q_\phi \| p_\nu) = \sum_{i=1}^{d} D_{\mathrm{KL}}^{(i)}$, with each term collecting into a closed-form scale term, a Student-t tail penalty, and a ν-dependent normaliser:

$$D_{\mathrm{KL}}^{(i)} = \underbrace{\tfrac{1}{2}\log\frac{\sigma_{c,i}^2}{\sigma_{\phi,i}^2}}_{\text{scale}} + \underbrace{\tfrac{\nu+1}{2}\,\mathbb{E}_{q_i}\big[\log\big(1 + (z-\mu_{c,i})^2/(\nu\sigma_{c,i}^2)\big)\big]}_{\text{Student-}t\text{ tail penalty}} + \underbrace{\zeta(\nu)}_{\text{normaliser}}, \tag{23}$$

where $\zeta(\nu) = \log\frac{\Gamma(\nu/2)}{\Gamma((\nu+1)/2)} + \frac{1}{2}\log\frac{\nu}{2} - \frac{1}{2}$ (so $\zeta(4) \approx -0.438$ nats). Only the tail-penalty expectation lacks a closed form; it is taken under the Gaussian q_i and estimated by a single

reparameterised Monte Carlo draw during training [24]. As a consistency check, letting $\nu \to \infty$ sends the tail penalty to $(\delta_i^2 + \sigma_{\phi,i}^2)/(2\sigma_{c,i}^2)$ with $\delta_i = \mu_{\phi,i} - \mu_{c,i}$, and (23) reduces exactly to the Gaussian–Gaussian KL.

To prevent posterior collapse we apply free bits, $D_{\mathrm{KL}} = \sum_{i=1}^{d} \max(\lambda_{\mathrm{free}}, D_{\mathrm{KL}}^{(i)})$ with $\lambda_{\mathrm{free}} = 0.25$ nats; for $d = 24$ latent dimensions this floors the KL at 6.0 nats, the value at which training stabilises (Section 4.4). This KL structure delivers the three properties stress testing requires: fat-tailed latent codes that make severe scenarios reachable under bounded KL, controlled tail penalisation that lets the decoder learn meaningful mappings for extreme $\mathbf{z}$, and crisis-severity alignment, since the prior parameters $(\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c)$ are learned functions of $\mathbf{c}$.

B Local-Volatility Pricing: Numerical Details

This appendix gives the full numerical pipeline that maps each generated implied-volatility surface to path-dependent exotic prices (summarised in Section 4.7).

Step 1: arbitrage-free SVI fit. Each generated surface $\hat{\boldsymbol{\sigma}}$ is post-processed via the SSVI fit of Section 3.3.3, yielding a continuous total-variance function $w(k, T)$ free of calendar and butterfly arbitrage.

Step 2: Dupire local volatility. The local-volatility function is recovered via the Dupire formula [37]:

$$\sigma_{\mathrm{loc}}^2(K, T) = \frac{\partial_T w(k, T)}{1 - \frac{k}{w}\partial_k w + \frac{1}{4}\left(-\frac{1}{4} - \frac{1}{w} + \frac{k^2}{w^2}\right)(\partial_k w)^2 + \frac{1}{2}\partial_{kk} w} \tag{24}$$

where $w(k, T) = \hat{\sigma}_{\mathrm{imp}}^2(k, T)\, T$ is the total implied variance from the SSVI fit, $k = \log(K/F_T)$ is the log-moneyness against the forward F_T, and $\partial_T w$, $\partial_k w$, $\partial_{kk} w$ denote the first calendar derivative, the first and second strike-slope derivatives of total variance, respectively. The SSVI parameterisation guarantees that the denominator (the Dupire "g-function") is strictly positive, this is precisely the no-butterfly condition, so σ_{loc}^2 is well-defined and non-negative throughout the strike–maturity domain.

Step 3a: PDE pricing (Crank–Nicolson). The barrier and no-touch options are priced by solving the local-volatility Black–Scholes PDE

$$\partial_t V + \tfrac{1}{2}\sigma_{\mathrm{loc}}^2(S, t)\, S^2\, \partial_{SS} V + r\, S\, \partial_S V - r\, V = 0, \tag{25}$$

with terminal condition $V(S, T) = (K - S)^+$ for the put and absorbing boundary $V(B, t) = 0$ for the knock-out, $V(B, t) = V_{\text{vanilla put}}(B, t)$ for the down-and-in case. We use a Crank–Nicolson scheme [38] with $n_S = 400$ spatial nodes log-spaced over $S \in [0.4\, S_0, 2.5\, S_0]$ and $n_t = 500$ time steps. The barrier is enforced via Andersen–Brotherton-Ratcliffe boundary tracking to recover second-order convergence.

Step 3b: Monte Carlo cross-check (Brownian bridge). As an independent verification, we price the same instruments by Monte Carlo simulation of the local-volatility SDE

$$dS_t = r\, S_t\, dt + \sigma_{\mathrm{loc}}(S_t, t)\, S_t\, dW_t, \tag{26}$$

discretised on $n_t = 252$ daily steps with $N_{\mathrm{MC}} = 100\,000$ paths. Discrete monitoring with daily steps introduces well-known positive bias for continuously-monitored barriers; we correct this with the Beaglehole–Dybvig–Zhou [39] Brownian-bridge formula, which adjusts the survival

probability per step by the analytic probability that the bridge between consecutive grid points crosses the barrier:

$$p_{\text{cross}}^{(i)} = \exp\left(-\frac{2(B - S_{t_i})(B - S_{t_{i+1}})}{\sigma_{\text{loc}}^2(S_{t_i}, t_i)\, S_{t_i}^2\, (t_{i+1} - t_i)} \right) \quad \text{if } S_{t_i}, S_{t_{i+1}} > B. \tag{27}$$

Antithetic variates and control variates (the corresponding vanilla put, priced analytically) reduce variance by approximately 4× versus naive MC.

C Bates Model Jump Parameters

The jump parameters used to generate the Bates robustness surfaces (Section 4.1) are calibrated by market regime as follows.

Table 19: Bates Jump Parameters by Market Regime

Regime	λ_J (jumps/yr)	μ_J (mean log-jump)	σ_J (jump vol)	Skew contribution
Normal	$[0.1, 0.5]$	$[-0.03, -0.01]$	$[0.02, 0.05]$	+3–5 pp
Stressed	$[0.5, 1.5]$	$[-0.06, -0.02]$	$[0.05, 0.10]$	+8–15 pp
Crisis	$[1.0, 3.0]$	$[-0.10, -0.03]$	$[0.08, 0.15]$	+15–25 pp

D Ablation Studies

The three architectural choices that most distinguish the Crisis VAE from a vanilla conditional VAE are (i) the crisis bottleneck dimension, (ii) the specific components of the crisis vector $\mathbf{c}$, and (iii) the Student-t prior degree of freedom ν. We ablate each in turn.

D.1 Bottleneck Dimension Ablation

The crisis bottleneck $\mathbf{c} \to \tilde{\mathbf{c}}$ compresses the 5-dimensional crisis indicator before it conditions the decoder, on the rationale that compression forces the latent code $\mathbf{z}$ to carry complementary information. We sweep the bottleneck dimension $d_{\tilde{c}} \in \{0, 1, 2, 3, 4, 5\}$, where $d_{\tilde{c}} = 5$ corresponds to no compression and $d_{\tilde{c}} = 0$ removes crisis conditioning from the decoder entirely (leaving only the prior network).

Table 20: Bottleneck Dimension Ablation: Architectural Behaviour and Stress Test Outputs

Config	$d_{\tilde{c}}$	Recon MSE ($\times 10^{-4}$)	KL nats	Dec. sens.	Enc. inform.	P&L std	VaR_{95}
No conditioning	0	9.8	9.4	0.18	0.09	2.34	4.85
Severe compression	1	4.2	7.8	0.09	0.04	0.96	7.12
Current	2	1.4	6.0	0.042	0.011	0.14	8.14
Mild compression	3	1.2	5.4	0.028	0.008	0.09	8.21
Minimal compression	4	1.1	4.9	0.019	0.005	0.06	8.27
No compression	5	1.0	4.6	0.014	0.003	0.04	8.31

Decoder sensitivity = $\max |\hat{x}(z_{\text{enc}}) - \hat{x}(z_{\text{rand}})|$; encoder informativeness = $\max |\hat{x}(z_{\text{enc}}) - \hat{x}(\mathbf{0})|$. Barrier P&L std for Crisis VAE Severe. All other hyperparameters held fixed.
The P&L std and VaR_{95} columns here and in Table 21 come from the earlier single-volatility pricer and are reported for *relative* architectural comparison; absolute magnitudes differ from the local-volatility headline (Table 11). The monotone effect of the bottleneck dimension is unaffected by the pricing route.

The pattern across the sweep is monotone and architecturally interpretable. With no conditioning ($d_{\tilde{c}} = 0$), the decoder relies on $\mathbf{z}$ alone, latent informativeness is highest (encoder

informativeness 0.09), but ATM vol control collapses, the barrier P&L standard deviation balloons to \$2.34, and the VaR_{95} drops to \$4.85 because crisis severity is no longer reliably targeted. At the other extreme ($d_{\tilde{c}} = 5$, no compression), the decoder reads the full crisis vector with no incentive to use $\mathbf{z}$; the latent code becomes informationally inert (informativeness 0.003) and within-crisis dispersion shrinks further (P&L std \$0.04). The current $d_{\tilde{c}} = 2$ choice sits at a deliberately middle point: it preserves enough $\mathbf{z}$ utilisation to support skew-sensitive instruments (Section 5.9) while delivering tight regulatory targeting. Moving to $d_{\tilde{c}} = 3$ or $d_{\tilde{c}} = 4$ further tightens the level control at the cost of within-regime diversity, a defensible alternative for jurisdictions whose stress-testing standards prioritise reproducibility above all else.

D.2 Crisis Component Ablation

The five-dimensional crisis vector $\mathbf{c} = (c_1, c_2, c_3, c_4, c_5)$ captures level, vol-of-vol, return proxy, skew, and term-structure slope. We retrain the model with each component held out (replaced by a constant zero) and re-evaluate at the severe-crisis specification.

Table 21: Crisis Component Ablation: Effect on Stressed-Surface Statistics

Held out	ATM 1Y vol (%)	RR 1M (%)	Term slope (pp)	Barrier P&L std	Mostly affected metric
None (full $\mathbf{c}$)	35.8 ± 0.4	13.2 ± 1.8	$+8.6 \pm 1.4$	0.14	,
c_1 (level)	26.4 ± 4.1	12.8 ± 1.7	$+8.4 \pm 1.5$	1.92	ATM vol level
c_2 (vol-of-vol)	35.6 ± 0.5	13.1 ± 1.9	$+8.5 \pm 1.5$	0.16	Marginal
c_3 (return proxy)	35.7 ± 0.4	12.5 ± 2.4	$+8.6 \pm 1.4$	0.21	Wing dispersion
c_4 (skew)	35.8 ± 0.4	4.8 ± 3.2	$+8.5 \pm 1.5$	0.18	Risk reversal collapse
c_5 (term slope)	35.8 ± 0.4	13.0 ± 1.7	$+1.2 \pm 4.8$	0.17	Term structure shape

The ablation cleanly identifies the role of each component. Removing c_1 destroys level control (ATM vol mean drops from 35.6% to 26.4% and std grows ten-fold). Removing c_4 collapses the risk reversal from 13.2% to 4.8%, confirming that c_4 is essential for generating crisis-consistent skews. Removing c_5 destroys term-structure control. The minor sensitivity of the level and overall risk-measures to c_2 removal suggests that vol-of-vol information is largely redundant with c_1 for surfaces in the training distribution, a candidate for principled dimensionality reduction in future architectures.

D.3 Student-t Degree-of-Freedom Sensitivity

Table 22: Student-t Degree-of-Freedom Sensitivity

ν	Recon MSE ($\times 10^{-4}$)	KL nats	Tail mass $P(\lvert z\rvert > 3)$	VaR_{99}	Stability
3	1.6	5.8	0.062	8.05	Occasional NaNs
4 **(ref)**	1.4	6.0	0.040	7.97	Stable
6	1.3	6.2	0.022	7.93	Stable
10	1.3	6.3	0.010	7.89	Stable
∞ (Gaussian)	1.3	6.3	0.003	7.84	Stable

The reference choice $\nu = 4$ matches the empirical tail index documented for equity-index returns ([11] reports $\nu \in [3, 5]$ across major indices), and is robustly trainable. Lighter-tailed priors ($\nu \to \infty$) lose the ability to generate the extreme latent codes required for severe stress scenarios, visible in the small but systematic drop in VaR_{99} from \$7.97 (Student-t_4) to \$7.84 (Gaussian). Heavier-tailed priors ($\nu = 3$) cause occasional training instability through the reparameterisation gradient's $1/\nu$ scaling near the mode.

D.4 Disentanglement Diagnostics: Linear Probes and Mutual Information

A bottleneck-conditioned VAE is only useful if the latent code $\mathbf{z}$ encodes information genuinely complementary to $\tilde{\mathbf{c}}$, rather than redundantly re-encoding crisis severity. We assess this via two standard disentanglement diagnostics [40, 41].

Linear probes. We train a linear regressor from $\mathbf{z}$ to each crisis component c_j on a held-out set of 3,000 surfaces and report the R^2. Low R^2 indicates that the corresponding crisis information has been successfully routed through $\tilde{\mathbf{c}}$ rather than $\mathbf{z}$.

Table 23: Linear-Probe R^2: Predicting Crisis Components from Latent Code $\mathbf{z}$

Target	c_1 (level)	c_2 (volvol)	c_3 (return)	c_4 (skew)	c_5 (term)
$R^2(\mathbf{z} \to c_j)$	0.18	0.14	0.09	0.31	0.27
$R^2(\tilde{\mathbf{c}} \to c_j)$	0.96	0.81	0.43	0.62	0.71
R^2(vanilla CVAE: $\mathbf{z} \to c_j$)	0.92	0.85	0.78	0.89	0.83

The crisis level (c_1) is overwhelmingly captured by $\tilde{\mathbf{c}}$ ($R^2 = 0.96$) and only weakly recoverable from $\mathbf{z}$ ($R^2 = 0.18$), a clean separation. The two components with the highest residual $\mathbf{z}$-content are c_4 (skew, $R^2 = 0.31$) and c_5 (term slope, $R^2 = 0.27$), consistent with the design intent that $\mathbf{z}$ captures surface-shape detail that the 2-dimensional bottleneck cannot fully transmit. The third row reports the same probes on a vanilla conditional VAE trained with $d_{\tilde{c}} = 5$ (no compression): every R^2 jumps above 0.78, confirming that without the bottleneck, the latent code redundantly re-encodes the crisis information.

Mutual information estimates. We estimate $I(\mathbf{z}; c_j)$ via the InfoNCE bound [41], averaged over 1,000 mini-batches. The estimates corroborate the linear-probe findings: $I(\mathbf{z}; c_1) \approx 0.18$ nats (small), $I(\mathbf{z}; c_4) \approx 0.45$ nats (moderate), $I(\tilde{\mathbf{c}}; c_1) \approx 2.84$ nats (large). The total information into the decoder is approximately $I(\tilde{\mathbf{c}}; \mathbf{c}) + I(\mathbf{z}; \mathbf{x}|\mathbf{c}) \approx 4.1 + 5.9 = 10.0$ nats, with the bottleneck carrying about 40% and the latent code about 60% of the decoder's effective input information.

E Multi-Confidence VaR/ES: Variance Swap and TARF

These tables complement the barrier multi-confidence results (Table 14) with the variance-swap and TARF analogues.

Table 24: Value-at-Risk and Expected Shortfall at Multiple Confidence Levels, Variance Swap

Method	VaR_{90}	ES_{90}	VaR_{95}	ES_{95}	$\text{VaR}_{97.5}{}^{\dagger}$	$\text{ES}_{97.5}{}^{\dagger}$	VaR_{99}	ES_{99}
Crisis VAE (Severe)	0.016	0.017	0.017	0.017	0.017	0.017	0.018	0.018
Crisis VAE (Moderate)	0.010	0.011	0.011	0.011	0.011	0.011	0.012	0.012
Crisis VAE (Mild)	0.005	0.005	0.005	0.005	0.005	0.005	0.006	0.006
Historical Simulation	0.025	0.027	0.027	0.028	0.028	0.028	0.029	0.029
Parametric (Δ-Γ)	0.008	0.011	0.010	0.013	0.013	0.014	0.016	0.018
GARCH(1,1) capped	0.022	0.024	0.024	0.024	0.024	0.024	0.024	0.024

GARCH VaR_{99} (\$0.024) and ES_{99} (\$0.024) exceed Crisis VAE (\$0.018, \$0.018) by roughly one third, driven by the heavier GARCH vol distribution even after capping.
VaR/ES at 95% and 97.5% and ES_{99} are computed directly; the 90% and 99% VaR columns are interpolated from the empirical quantiles.
† FRTB IMA confidence level [42]; interpolated between empirical 95% and 99% quantiles.

Table 25: Value-at-Risk and Expected Shortfall at Multiple Confidence Levels, TARF

Method	VaR_{90}	ES_{90}	VaR_{95}	ES_{95}	$\text{VaR}_{97.5}^{\dagger}$	$\text{ES}_{97.5}^{\dagger}$	VaR_{99}	ES_{99}
Crisis VAE (Severe)	0.045	0.044	0.045	0.044	0.045	0.044	0.044	0.043
Crisis VAE (Moderate)	0.041	0.040	0.040	0.040	0.040	0.040	0.040	0.039
Crisis VAE (Mild)	0.033	0.032	0.032	0.032	0.032	0.032	0.032	0.031
Historical Simulation	0.038	0.037	0.037	0.037	0.037	0.037	0.036	0.036
Parametric (Δ-Γ)	-0.001	-0.013	-0.004	-0.019	-0.014	-0.020	-0.026	-0.046
GARCH(1,1) capped	0.039	0.036	0.037	0.035	0.035	0.035	0.034	0.032

TARF shows low stress sensitivity due to its knockout structure (cumulative gain cap of 15), which triggers early under any substantial vol increase.

† FRTB IMA confidence level [42]; interpolated between empirical 95% and 99% quantiles.

F Qualitative Method Comparison

Table 26: Qualitative Feature Comparison Across Methods

Feature	Crisis VAE	Historical Sim	Parametric	GARCH(1,1)	Expert
Surface consistency	Full surface	Full surface	Parallel only	Parallel only	Parallel only
Smile dynamics	Learned	Historical	None	None	None
Novel scenarios	Yes (latent)	No (resample)	Yes (Gaussian)	Yes (GARCH)	No (fixed)
Crisis conditioning	Explicit ($\mathbf{c}$)	Implicit	None	Implicit	Explicit
Arbitrage-aware	Soft penalty	From data	No guarantee	No guarantee	No guarantee
Tail capture	Student-t prior	Data-limited	Normal tails	Capped GARCH	None
Distributional output	Yes (N samples)	Yes (N samples)	Yes (analytic)	Yes (N paths)	No (1 point)
Sign reversal risk	No	No	Yes (barrier)	No	No

The Crisis VAE is the only method that simultaneously provides: (i) complete volatility surfaces with learned smile dynamics, (ii) explicit crisis severity control, (iii) distributional risk output, and (iv) no sign reversal risk for barrier options.

G Bates Robustness Details

The Bates surfaces demonstrate that enriching the training distribution with jump dynamics does not alter the key surface properties at the 90%–110% moneyness range. Risk reversals at this range show remarkable stability: Normal 9.0% (Heston) vs. 8.9% (Bates), Stressed 16.8% vs. 14.1%, Crisis 22.4% vs. 22.5%. The Parametric sign reversal arises from the payoff structure, not the volatility model, it persists regardless of whether synthetic surfaces come from Heston or Bates data, a finding that follows from the mathematical structure of the barrier payoff independently of the surface generator. GARCH residual tail uncertainty similarly persists under alternative training distributions. The method ranking (Crisis VAE most conservative and tightest; Parametric sign reversal; GARCH residual tail) is therefore invariant to the choice of synthetic training distribution. Extension to real market surfaces remains an important direction for future empirical validation.

Table 27: Bates Surface Statistics (5,000 Surfaces)

Regime	ATM 1M (%)	ATM 2Y (%)	RR 1M (%)	RR 1Y (%)
Normal	17.4 ± 2.3	23.0 ± 2.9	8.9 ± 2.1	10.0 ± 0.6
Stressed	27.3 ± 2.5	38.3 ± 5.0	14.1 ± 7.8	13.2 ± 1.0
Crisis	37.1 ± 3.3	61.2 ± 9.2	22.5 ± 2.3	17.2 ± 1.2

G.0.1 Surface Shape Statistics Under Severe Crisis

Table 28: Surface Shape Statistics: 500 Severe-Crisis VAE Samples

Metric	Mean ± Std	Range
ATM 1Y IV (%)	35.8 ± 0.4	[33.7, 36.6]
Risk reversal 1M (%)[a]	surface-dependent	regime-controlled
Term structure slope (pp)[b]	controlled by $\tilde{\mathbf{c}}$	via bottleneck
Calendar spread violations (%)	< 0.8	,
Butterfly convexity compliance (%)	> 97	,

[a] IV(90% moneyness) − IV(110% moneyness) at short tenor.
[b] ATM 2Y IV − ATM 1M IV.

While ATM vol is tightly controlled (±0.4%), the surface *shape* exhibits variation driven by $\mathbf{z}$, skew shape, curvature patterns, and wing behavior, confirming the intended separation of concerns in the architecture. The narrow P&L dispersion for the barrier option arises because this instrument is primarily sensitive to ATM vol level rather than skew or curvature, a property of the instrument, not a deficiency of the model.

Compliance with Ethical Standards

- **Clinical Trial Number:** Not applicable
- **Trial Registration:** Not applicable
- **Consent to Publish declaration:** Not applicable
- **Consent to Participate declaration:** Not applicable
- **Ethics declaration:** Not applicable

References

[1] Acerbi, C. and Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503.

[2] Alexander, C. and Sheedy, E. (2008). Developing a stress testing framework based on market risk models. *Journal of Banking & Finance*, 32(10):2220–2236.

[3] Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228.

[4] Basu, S. (2008). Stress testing with volatility-weighted historical simulation. *Journal of Risk Management in Financial Institutions*, 1(4):345–368.

[5] Basel Committee on Banking Supervision (2009). Principles for sound stress testing practices and supervision. Bank for International Settlements.

[6] Basel Committee on Banking Supervision (2016). Minimum capital requirements for market risk. Bank for International Settlements.

[7] Bates, D. S. (1996). Jumps and stochastic volatility: Exchange rate processes implicit in Deutsche Mark options. *The Review of Financial Studies*, 9(1):69–107.

[8] Bergeron, M., Fung, N., Hull, J., and Poulos, Z. (2021). Variational autoencoders: A hands-off approach to volatility. *arXiv preprint arXiv:2102.03945*.

[9] Breeden, D. T. and Litzenberger, R. H. (1978). Prices of state-contingent claims implicit in option prices. *Journal of Business*, 51(4):621–651.

[10] Carr, P. and Madan, D. (1998). Towards a theory of volatility trading. In *Volatility*, pages 417–427. Risk Books.

[11] Cont, R. (2001). Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236.

[12] Demeterfi, K., Derman, E., Kamal, M., and Zou, J. (1999). A guide to volatility and variance swaps. *The Journal of Derivatives*, 6(4):9–32.

[13] Duffie, D. (2001). *Dynamic asset pricing theory*. Princeton University Press.

[14] Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events: for insurance and finance*. Springer Science & Business Media.

[15] Feugang Nteumagné, B., Azemtsa Donfack, H., and Wafo Soh, C. (2025). Variational autoencoders for completing the volatility surfaces. *Journal of Risk and Financial Management*, 18(5):239. `https://doi.org/10.3390/jrfm18050239`.

[16] Gatheral, J. (2006). *The volatility surface: a practitioner's guide*. John Wiley & Sons.

[17] Glasserman, P. (2003). *Monte Carlo Methods in Financial Engineering*. Springer.

[18] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

[19] Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

[20] Harrison, J. M. and Kreps, D. M. (1979). Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory*, 20(3):381–408.

[21] Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *The Review of Financial Studies*, 6(2):327–343.

[22] Hull, J. C. (2018). *Options, Futures, and Other Derivatives*. Pearson, 10th edition.

[23] Karatzas, I. and Shreve, S. E. (1991). *Brownian motion and stochastic calculus*. Springer Science & Business Media.

[24] Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

[25] Kondratyev, A. (2019). Market generators with GAN. *Risk*, 32(8):34–39.

[26] Kotz, S. and Nadarajah, S. (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.

[27] Merton, R. C. (1973). Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 4(1):141–183.

[28] Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1–2):125–144.

[29] Rebonato, R. (2010). *Coherent stress testing: A Bayesian approach to the analysis of financial stress*. John Wiley & Sons.

[30] Sohn, K., Lee, H., and Yan, X. (2015). Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, 28.

[31] Tishby, N., Pereira, F. C., and Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

[32] Wystup, U. (2006). *FX options and structured products*. John Wiley & Sons.

[33] Wiese, M., Knobloch, R., Korn, R., and Kretschmer, P. (2019). Quant GANs: Deep generation of financial time series. *Quantitative Finance*, 20(9):1419–1440.

[34] Kienitz, J. (2025). Gaussian GenAI: Synthetic market data generation with Gaussian mixture models. *SSRN Working Paper No. 5050372*.

[35] Hounwanou, C. D., Ntakirutimana, P., and Gaba, Y. U. (2025). Evaluating generative models for synthetic financial data. *arXiv preprint arXiv:2512.21791*.

[36] Gatheral, J. and Jacquier, A. (2014). Arbitrage-free SVI volatility surfaces. *Quantitative Finance*, 14(1):59–71.

[37] Dupire, B. (1994). Pricing with a smile. *Risk*, 7(1):18–20.

[38] Andersen, L. and Brotherton-Ratcliffe, R. (1998). The equity option volatility smile: an implicit finite-difference approach. *Journal of Computational Finance*, 1(2):5–37.

[39] Beaglehole, D. R., Dybvig, P. H., and Zhou, G. (1997). Going to extremes: correcting simulation bias in exotic option valuation. *Financial Analysts Journal*, 53(1):62–68.

[40] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). β-VAE: Learning basic visual concepts with a constrained variational framework. *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

[41] Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. (2018). Isolating sources of disentanglement in variational autoencoders. *Advances in Neural Information Processing Systems*, 31.

[42] Basel Committee on Banking Supervision (2019). Minimum capital requirements for market risk (revised). Bank for International Settlements, January 2019.

[43] Cont, R. and da Fonseca, J. (2002). Dynamics of implied volatility surfaces. *Quantitative Finance*, 2(1):45–60.

[44] Rouah, F. D. (2013). *The Heston model and its extensions in MATLAB and C#*. John Wiley & Sons.