GENERATIVE FINANCE PROJECT

Multi-Asset Scenario Stress Lab

Evidence-Constrained Scenario Analysis for Portfolio Stress Decision Support

Technical White Paper

System architecture, validation evidence, and operating boundaries · August 2026

Sungkyu Lee

Graduate School of Computing, Yonsei University

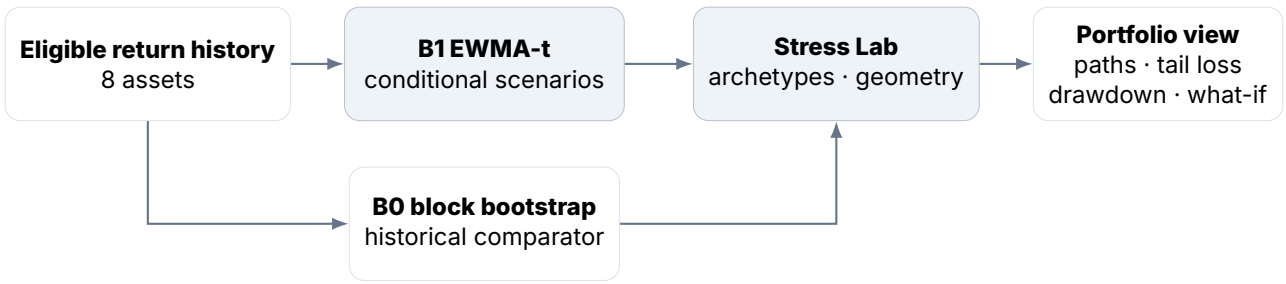
scott.lee@yonsei.ac.kr

ORCID: 0009-0008-5598-3346

What this system is. The Multi-Asset Scenario Stress Lab is a human-in-the-loop operational research system for 20-trading-day multi-asset stress analysis. It generates conditional joint return paths for eight liquid exchange-traded funds (ETFs), organizes adverse scenarios into transparent stress archetypes, compares those conditional scenarios with a historical block-bootstrap benchmark, and revalues portfolios on the same scenario cloud.

What the evidence supports. The pre-specified model-selection process and primary out-of-sample evaluation did *not* establish predictive superiority for the modern generative models. A later, explicitly retrospective decision-value evaluation supported only one practical use: the **B1 EWMA-t scenario-archetype Stress Lab**. The other tested uses did not meet their qualification criteria: predictive downside signaling, a dynamic defensive overlay, tail-driver attribution, and a B1-versus-Flow disagreement warning.

What this system is not. It is not a crash-prediction engine, trading signal, automatic allocation rule, or calibrated crisis-probability estimator. Nor does the reported evidence establish that a deep generative model outperforms the conventional B1 EWMA-t simulator.



8 ETFs · 20-day horizon · 5,000 scenarios · B1 vs historical · human-in-the-loop

System role: Scenario-Based Portfolio Stress Decision Support · Core model: B1 EWMA-t · Historical comparator: B0 block bootstrap
Flow F1: research challenger only

Abstract

The practical value of a multi-asset scenario generator depends on more than whether simulated paths look plausible. Its predictive distribution must be reasonably calibrated, its joint dependence structure must be credible, and each proposed practical use must satisfy pre-specified qualification criteria against an appropriate transparent benchmark. This white paper documents the evidence-constrained development of the **Multi-Asset Scenario Stress Lab**, an operational research system for scenario-based portfolio stress decision support.

The system covers eight liquid ETFs—SPY, EFA, EEM, IEF, LQD, GLD, DBC, and VNQ—and generates 5,000 joint 20-day return paths at each forecast origin. The 60-trading-day lookback applies only to the neural challengers, which use the trailing return tensor as direct conditioning information. B0 and B1 instead use all eligible history through the forecast origin: B0 resamples contiguous 20-day blocks, whereas B1 estimates a return mean and an exponentially weighted covariance matrix and draws heavy-tailed multivariate Student-t innovations. The evaluation design uses 2006–2014 for initial training, 2015–2018 for validation and model selection, and 2019–2025 for a single primary out-of-sample (OOS) evaluation with no OOS tuning. The candidate set includes historical block resampling, EWMA-t, DCC-GARCH and hidden-Markov benchmarks, a conditional variational autoencoder, a conditional diffusion model, and conditional Flow Matching.

The primary OOS evidence did not establish predictive superiority for the modern generative models. B1, an exponentially weighted moving-average Student-t (EWMA-t) simulator, achieved the best mean 20-day Energy Score and standardized continuous ranked probability score (CRPS), the lowest broad calibration errors, and lower errors on the main dependence diagnostics. Flow F1 remained competitive and much faster than the 40-step Flow reference, but it was not superior to B1. A later five-use-case decision-value program (Stages A, B-P, C, D, and E) tested proposed practical uses. Because the 2019–2025 interval had already been examined, this program is explicitly retrospective rather than a new independent OOS test. Four uses did not meet their pre-specified criteria. The only retained use was a transparent **scenario-archetype Stress Lab**: across 56 large-drawdown origins fixed for the retrospective evaluation, B1 placed more tail-scenario mass on the archetype matching the subsequently realized stress pattern and produced a representative cross-asset stress vector closer to the realized outcome than B0 historical resampling. It also passed the pre-specified simulation-stability and leave-one-year-out checks.

The resulting product is intentionally narrow. Stress-family shares are frequencies within the generated scenario tail, not calibrated real-world event probabilities. The system does not issue trade signals, automatic defensive allocations, model-confidence traffic lights, optimized portfolio weights, or model-switching rules. Its contribution is therefore not a new generative model. Instead, the white paper documents a system-design process for converting probabilistic scenario research into bounded decision support: test each intended use, keep a transparent comparator visible, and remove features that fail their qualification criteria.

Evidence hierarchy used in this paper. Model selection and primary OOS qualification are based on the pre-specified predictive-distribution evaluation. The later Stage A–E work is a *retrospective decision-value evaluation* conducted on the same 2019–2025 interval already used for primary OOS evaluation; the Stage B-P portfolio overlay is additionally a *post-OOS policy study*. These later stages must not be described as new independent OOS strategy tests.

Executive Summary

Research question	Can conditional multi-asset scenarios add decision-useful stress information beyond transparent historical resampling, and which practical uses meet pre-specified qualification criteria?
Core product	An operational Stress Lab centered on B1, with B0 kept visible as the historical comparator and portfolio what-if analysis performed on a common B1 scenario cloud.
Research universe	Eight liquid ETFs: SPY, EFA, EEM, IEF, LQD, GLD, DBC, and VNQ. BIL, a short-term U.S. Treasury-bill ETF, is excluded from scenario generation and appears only in the retrospective B-P policy study.
Evaluation setup	20-day forecast horizon; forecast origins every five trading days; 5,000 joint scenarios per OOS origin. The neural challengers use a 60-day conditioning window, whereas B0 and B1 use all eligible history through the forecast origin.
Primary OOS	348 fixed forecast origins from 2019-01-02 through 2025-11-25. Model and hyperparameter choices were not changed in response to the OOS results.
Model conclusion	B1 EWMA-t remains the reference model. Flow F1 remains an OOS-competitive, research-only challenger. The evidence does not support predictive superiority for any modern generative model.
Use-case result	A FAIL / B-P FAIL / C FAIL / **D PASS** / E FAIL. Only the Stage-D stress-archetype use is retained in the operational product.
Public boundary	Generated-scenario frequencies are descriptive model outputs, not calibrated event probabilities. The product exposes no automated trading, predictive risk traffic light, automatic BIL overlay, or model switching/blending.

Evidence-to-Design Consequences. The later use-case evaluation narrowed the product rather than expanding it. Stage A did not qualify a predictive risk alert. Stage B-P did not qualify an automated defensive allocation or BIL overlay. Stage C did not qualify a tail-driver attribution layer. **Stage D alone qualified** the bounded stress-archetype comparison against B0 historical resampling. Stage E did not qualify a model-disagreement confidence warning. These dispositions define the current operational scope.

Contents

1 Why a Scenario Stress Lab?

1.1 The decision problem is distributional, path-dependent, and joint

A point forecast answers a relatively narrow question such as "what is the expected outcome?" Portfolio stress decisions require more. A practitioner needs to know what range of adverse outcomes is plausible, how assets may move together in those outcomes, and how losses can develop before the horizon ends. Three features therefore matter simultaneously:

- **Distributional uncertainty:** the system must represent a range of possible outcomes and their tails rather than a single expected return.
- **Joint dependence:** the scenarios must preserve cross-asset co-movement, because diversification can weaken precisely when several markets sell off together.
- **Path dependence:** the order and depth of daily moves matter. Two paths can finish with the same 20-day return yet produce very different maximum drawdowns or worst-five-day losses.

The Stress Lab represents one future scenario as a matrix of returns

$$\mathbf{R}^{(s)} \in \mathbb{R}^{H \times A}, \qquad H = 20,\ A = 8,$$

where each row is a future trading day and each column is one of the eight assets. Scenario s therefore contains a complete 20-trading-day joint path, not eight unrelated asset forecasts. At each forecast origin the system samples $N = 5{,}000$ such paths. A portfolio path is then computed from the same set of joint scenarios by applying portfolio weights to each simulated day, so alternative portfolios can be compared without regenerating a different market scenario set for each one.

> **Core design principle.** The object being evaluated is the *scenario distribution as a whole*, not whether a few individual simulated paths look convincing. A credible model should meet its calibration tests. It should also preserve relevant cross-asset and path behavior, remain stable across Monte Carlo draws, and add useful information beyond simpler comparators.

1.2 A system-first research architecture

The operational research system is the primary maintained product. The Full Manual, this Technical White Paper, an SSRN archival version, validation records, source code, and release notes each serve a distinct supporting role. The approach is intentionally system-first rather than paper-first: the dashboard is not a visual demonstration built around a manuscript. It exists because one narrow practical use met the qualification criteria. This white paper documents the design, the supporting evidence, the rejected uses, and the limits of what the system is validated to do.

Product scope is therefore an empirical outcome rather than a design preference. The research program examined a broader set of predictive and practical functions, but the operational product retains only those uses supported by the reported evidence. Features that were attractive in concept but failed their pre-specified criteria were removed from the public operational scope.

1.3 Research questions

The project can be summarized by four questions:

1. Can conditional or deep generative scenario models improve the quality of a multi-asset predictive distribution relative to transparent conventional baselines?
2. Do scenario-based tail summaries forecast subsequent realized downside strongly enough to function as a risk signal?
3. If predictive signaling fails, do the scenarios still represent realized stress structure more usefully than historical resampling for human portfolio review?
4. Which operational outputs can be exposed publicly without turning descriptive scenario shares into unsupported probability, timing, or trading claims?

2 Data, Information Clock, and Scenario Contract

2.1 Eight-asset research universe

The core universe is intentionally small and economically interpretable. It spans major equity, U.S. Treasury duration, investment-grade credit, and real-asset exposures, which is enough to create meaningful cross-asset stress patterns while keeping the scenario taxonomy auditable. The universe was not selected to maximize portfolio performance.

Table 1: Eight-asset scenario universe

Ticker	Economic role
SPY	U.S. equity
EFA	Developed equity ex-U.S.
EEM	Emerging-market equity
IEF	Intermediate U.S. Treasury
LQD	U.S. investment-grade corporate credit
GLD	Gold
DBC	Broad commodities
VNQ	U.S. real-estate investment trusts (REITs)

Research sample. The reported research evidence uses the Center for Research in Security Prices (CRSP) CIZ Flat File Format 2.0 daily total-return field (`DlyRet`), accessed through Wharton Research Data Services (WRDS). Simple total returns are converted to daily log total returns for model estimation and scenario generation; log returns are convenient for multivariate path modeling and are converted back to simple returns when portfolio values are calculated. After aligning the eight ETFs on common trading dates, the panel contains 5,008 daily observations from 2006-02-06 through 2025-12-31. Initial training uses the common start through 2014-12-31, validation and model selection use 2015–2018, and the primary OOS evaluation uses 2019–2025. The last OOS origin is 2025-11-25 because a 20-trading-day forecast starting any later would extend beyond the fixed 2025 research sample. This panel underlies all research results reported in the white paper.

Operational continuation. Current on-demand runs use Yahoo Finance adjusted daily prices after the CRSP research sample ends. A 2025 overlap comparison showed near-identical daily returns and met the continuity criteria for all eight ETFs. The bridge allows the operational dashboard to continue updating without rewriting the historical research evidence. It does *not* mean that Yahoo and CRSP are literally the same data source, does not change the reported OOS results, and does not create an exact historical point-in-time reconstruction of Yahoo data.

2.2 Information clock

A forecast origin t is defined immediately after the U.S. market close. At that point, returns through close t are observable and may be used; the next 20 close-to-close returns, $t+1, \ldots, t+20$, are still future outcomes and form the evaluation target.

$$\text{close } t \to \mathcal{I}_t \ \left(\text{neural input } X_{t-59:t}\right) \to \text{scenario generation} \to Y_{t+1:t+20},$$

where $\mathcal{I}_t$ denotes the information available through close t. This timing rule keeps future data out of both model estimation and evaluation. The implementation prohibits future closes and centered statistics. It also prohibits normalization with statistics from the future target window and any hyperparameter change made after primary-OOS results have been inspected. The model family uses market returns only; no discretionary macro regime label is added. The neural challengers take the trailing 60-day return tensor as their direct input. B0 and B1 are different: they may use the full eligible return history through t for historical block resampling and EWMA estimation, respectively. The 60-day neural lookback is therefore not a common estimation window for every model.

2.3 Scenario contract

Every model must produce the same output schema—the same dimensions, horizon, asset order, and log-return units—even though the models use different estimation and sampling mechanisms. Given the information available at the forecast origin, the fitted model state, and a random-number-generator (RNG) seed,

$$\text{eligible history} + \text{model state} + \text{RNG seed} \longrightarrow N \times H \times A \text{ return paths},$$

where N is the number of scenarios, H is the forecast horizon, and A is the number of assets. In the standard configuration, $N = 5{,}000$, $H = 20$, and $A = 8$, so every model returns a $5{,}000 \times 20 \times 8$ scenario cube. The RNG seed fixes the simulation randomness for a given model state and provides a reproducibility handle for the sampled scenario cloud under the same implementation environment. The contract standardizes this downstream path object—not the models' internal estimation histories—so the same stress analytics, portfolio calculations, audit records, and user interface can be applied to historical resampling, parametric simulation, diffusion, or Flow Matching.

The common evaluation and operational configuration is:

- neural-model conditioning input: the trailing 60 trading days of eight-asset returns;
- B0 resampling pool and B1 estimation sample: all permissible returns through the forecast origin;
- forecast horizon: 20 trading days;
- primary OOS forecast-origin stride: five trading days;
- final scenario count: 5,000 paths per origin; a pre-OOS simulation-stability check found that 1,000 paths produced overly noisy reported summaries;
- portfolio calculation convention: modeled log returns are converted to simple returns, and fixed target weights are applied on each simulated day, equivalent to constant-weight daily rebalancing over the 20-day analytical horizon.

The last convention is an evaluation device for consistent path-by-path portfolio comparison; it is not a recommendation to rebalance a live portfolio every day.

3 Model Families and Assigned Roles

The candidate set ranged from simple historical and parametric baselines to richer conditional generators. Besides B0 and B1, the conventional comparison included DCC-GARCH (dynamic conditional correlation GARCH) [4] and hidden Markov model (HMM) variants. DCC-GARCH lets conditional cross-asset correlations evolve with volatility; the HMM variants represent a small number of latent market regimes. On the 2015–2018 validation sample, B1 had the lowest mean Energy Score, Variogram Score, and standardized CRPS among B0, B1, DCC-GARCH, and HMM. It therefore became the conventional reference before the primary OOS comparison with the modern generators. The later OOS and use-case evidence then determined which models, if any, received an operational role.

Current roles after the full evidence review. B0 = historical stress comparator; B1 EWMA-t = reference model and Stress-Lab core; Flow F1 = research challenger only. DCC-GARCH, HMM, CVAE, DDPM, and Flow F3 remain documented research alternatives rather than part of the core operational model set.

3.1 B0: moving-block historical resampling

B0 draws contiguous 20-day return blocks from all permissible history available through the forecast origin, following moving-block bootstrap logic for dependent observations [3]. Because each sampled block is an actual historical sequence, B0 preserves the local serial pattern and cross-asset co-movement that occurred inside that block without fitting a conditional covariance model. This makes the method transparent and historically grounded.

Its limitation is equally clear: block selection is not conditioned on the market state at the current forecast origin. A historical episode can therefore be sampled even when its volatility and dependence structure differ from the current environment, and stress patterns with few historical analogs may be sparsely represented. B0 is consequently retained as the **historical stress comparator**, not as the official conditional engine.

3.2 B1: EWMA-t conditional simulator

B1 is an exponentially weighted moving-average Student-t simulator (EWMA-t). It uses all eligible returns through the forecast origin to estimate the return mean and an exponentially weighted covariance matrix, then draws innovations from a multivariate heavy-tailed Student-t distribution. Validation selected

$$\lambda = 0.99, \qquad \nu = 8,$$

where λ controls how slowly past covariance information decays and ν controls the heaviness of the Student-t tails. With $\lambda = 0.99$, recent observations receive more weight, but older eligible history is not discarded at a 60-day cutoff.

At a given forecast origin, B1 holds the estimated mean and covariance fixed across the 20 simulated days and draws independent multivariate Student-t daily innovations. In practical terms, it represents the current covariance environment and allows unusually large joint shocks, while remaining much simpler than a model whose volatility or regime state evolves inside each simulated path. B1 therefore does *not* model a changing latent volatility or regime over the 20-day horizon. It also has very low inference cost and no neural-network training loop.

B1 is the core model because of the evidence, not because it is simpler. It was the strongest conventional model on validation, provided the most reliable overall calibration/dependence anchor in primary OOS, and later met the Stage-D Stress-Lab criteria relative to B0.

3.3 Modern generative challengers

The modern challengers all solve the same conditional task—a 60×8 history tensor to a distribution over 20×8 future paths—but represent that conditional distribution in different ways.

CVAE. The conditional variational autoencoder (CVAE) [5] compresses the observed history into a context representation, samples a lower-dimensional latent variable, and decodes the context and latent draw into a future multi-asset path. This provides a relatively simple deep-generative benchmark. It did not qualify at validation: its mean Variogram Score, which is sensitive to joint dependence, was about 20.5% worse than B1.

DDPM. The denoising diffusion probabilistic model (DDPM) [6] starts from a noisy future-path tensor and learns to iteratively denoise it while conditioning on the 60-day history. Its validation quality was competitive with B1, but paired block-bootstrap diagnostics did not establish a decisive score advantage. It remained a research model after primary OOS rather than entering the core operational system.

Flow Matching. Conditional Flow Matching [7] starts from a simple base distribution and learns a continuous history-conditioned vector field that transports those samples toward the distribution of future paths. The DDPM and Flow implementations share the same temporal conditioning backbone and closely matched capacity, so the comparison is intended to reflect the generative objective and sampler rather than a larger network. Flow became the main modern research challenger because it combined competitive distributional quality with a favorable latency frontier, but it still did not establish predictive superiority over B1.

Three Flow solver budgets were evaluated before OOS using the same trained network: F1 with 10 integration steps, F2 with 20, and F3 with 40. The original pre-specified validation rule, which prioritized mean Energy Score, selected F3 as the higher-step research specification. A separate pre-OOS quality/latency audit then retained F1 as the efficiency configuration: its median validation latency was about 74% lower than F3, while the principal 20-day scores were nearly unchanged. Primary OOS evidence supported keeping F1 for research comparison, but not replacing B1.

> **No ensemble was added after OOS results were observed.** Constructing a B1+Flow blend at that point would have created a new model specification using knowledge of the evaluation results. The system therefore preserves the tested evidence as-is: B1 remains the reference model, while Flow F1 remains a research-only challenger.

4 Validation Architecture

4.1 Chronological separation

The research design separates development and model selection from a single primary out-of-sample (OOS) evaluation whose model choices were fixed before OOS results were inspected:

Initial estimation and annual expanding-window refits follow the pre-specified timing rule: an OOS year's model may use only data available through the prior year-end, while hyperparameters remain fixed after validation.

4.2 Proper scores, calibration, tails, and dependence

A scenario generator is a multivariate probabilistic forecast, not a point predictor. It must therefore be judged from several angles: whether the realized outcome is compatible with the forecast distribution, whether uncertainty bands have sensible coverage, whether portfolio tails and paths are represented adequately, and whether cross-asset dependence is credible. No single metric can answer all of those questions.

Diagnostic	Question	Examples used
Joint distribution	Does the joint 8-asset forecast assign low loss to the realized outcome?	Energy Score, Variogram Score
Marginal calibration	Do asset-level forecast intervals contain realized returns at approximately their stated rates?	CRPS, probability integral transform (PIT), 50/80/90/95% coverage
Portfolio tails	Do realized portfolio losses cross model-implied tail thresholds at approximately the nominal rates?	1/5/10% Value-at-Risk (VaR) breach rates, quantile (pinball) loss, Expected Shortfall (ES) diagnostics
Path risk	Do scenarios represent realized drawdown, volatility, and worst-window behavior?	Maximum drawdown, realized volatility, worst-5d return
Dependence	Do scenarios preserve co-movement and hedge relationships?	Correlation mean absolute error (MAE), downside co-exceedance, SPY–IEF correlation-sign Brier score
Simulation stability	Are reported summaries stable to scenario count and Monte Carlo randomness?	1,000-vs-5,000 comparison, split-half tests, seed diagnostics

How to read the main diagnostics. The proper scores are losses, so lower is better. CRPS evaluates a full univariate predictive distribution; standardized CRPS divides by a fixed validation-period return scale so assets with different volatilities are comparable. The Energy Score extends distributional scoring to the joint multi-asset outcome, while the Variogram Score adds sensitivity to pairwise differences and dependence [1,2]. PIT and interval coverage ask whether realized returns occupy the forecast distribution at the expected frequencies. VaR breach rates perform an analogous lower-tail check, ES summarizes average severity beyond a tail cutoff, and the Brier score is squared probability error for a binary event. These diagnostics are complementary; no single score determines a model's role.

4.3 Comparator discipline

Complexity is not itself evidence of value. A modern generator must first be credible relative to conventional distributional baselines, and each proposed practical use must then be compared with the simplest transparent method that can answer the same question. For the retained Stress-Lab use case, B0 historical block resampling can already produce stress-family frequencies and representative stress geometry without a conditional model. B1 therefore earns incremental value only if its conditional scenario representation improves those outputs relative to B0.

5 Primary Out-of-Sample Evidence

5.1 Primary OOS execution

The primary OOS evaluation used 348 forecast origins from 2019-01-02 through 2025-11-25 and 5,000 scenarios per origin. Model families, hyperparameters, refit timing, and evaluation definitions were fixed beforehand. An initial run halted because of an implementation fault before any valid OOS result was produced or viewed. After correcting only the execution and result-persistence layer, the full analysis was rerun under the unchanged scientific specification. The 2026 operational extension was excluded.

5.2 Joint-distribution results

Table 2: Primary 20-day OOS distributional scores (lower is better)

Model	Energy	Energy vs B1	Variogram	Std. CRPS
B1 EWMA-t	**1.4441**	–	3.4729	**0.4271**
Flow F1	1.4589	+1.03%	**3.3145**	0.4324
Flow F3	1.4606	+1.15%	3.3167	0.4329
DDPM	1.4872	+2.99%	3.3381	0.4413

"Energy vs B1" is the percentage change in mean Energy Score relative to B1; positive values are worse because lower scores are preferred.

Because origins are five trading days apart but each target spans 20 days, neighboring OOS outcomes overlap and score differences are serially dependent. The pre-specified 12-origin moving-block bootstrap therefore resamples short runs of adjacent origins rather than treating all 348 origins as independent. For both Energy and Variogram, every modern-model-minus-B1 95% interval included zero. The inference therefore does not establish modern-model superiority over B1 on either score. Flow's lower mean Variogram Score remains descriptive evidence, not an overall win.

5.3 Calibration and portfolio-tail asymmetry

A nominal 90% central interval should contain the realized return about 90% of the time if the distribution is well calibrated. Averaged across assets, mean absolute coverage error over the 80/90/95% intervals at 20 days was about 1.62 percentage points for B1 versus roughly 3.1 for each modern candidate. This supports treating generated frequencies as model outputs, not calibrated real-world event probabilities.

The portfolio tails show why that restriction matters. A nominal 5% VaR threshold should be breached about 5% of the time. Equal-weight breach frequencies were 4.31% for B1 and 5.17% for Flow F1; for the 60/40 portfolio they were 6.90% and 10.92%. Thus both models produced 60/40 breaches more often than the nominal rate, especially Flow. Calibration is portfolio-dependent, so no generated frequency is authorized as a universal probability.

5.4 Dependence and path behavior

B1 also had lower error on the broad dependence checks: pairwise correlation MAE was 0.219 versus 0.270 for Flow F1, and the Brier score for whether realized 20-day SPY–IEF correlation was positive was 0.204 versus 0.385. Flow F1 nevertheless had lower mean CRPS on maximum drawdown, realized volatility, and worst-five-day return for both equal-weight and 60/40 portfolios. Those path-risk differences lacked a pre-specified matched-pair inference test and remain descriptive, not superiority evidence. OOS latency also favored F1 over F3 (1.42 versus 5.75 seconds per 5,000 scenarios; B1 about 0.06 seconds), supporting F1 only as the efficient Flow research configuration.

> **Primary OOS outcome.** B1 EWMA-t remains the reference model. Flow F1 remains research-only: competitive, but not superior. DDPM and Flow F3 receive no operational role; CVAE failed validation. No modern generator supports a predictive-superiority claim over B1.

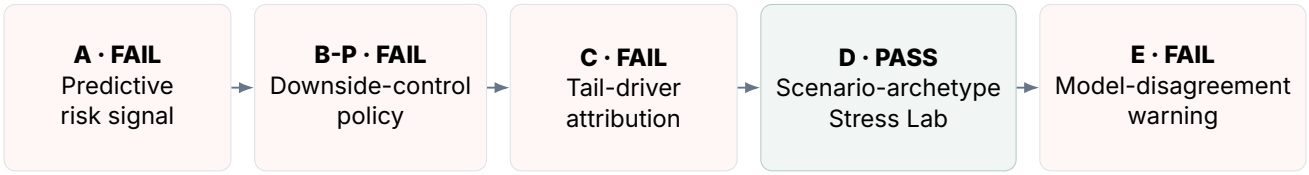
6 Retrospective Decision-Value Evaluation

6.1 Why a second validation layer was needed

A statistically respectable scenario model may still be unhelpful for a particular decision. Conversely, a distribution that fails as a market-timing signal may still help a practitioner describe the structure of adverse joint outcomes. The project therefore treated **model qualification** and **practical-use qualification** as separate questions.

The 2019–2025 primary OOS results had already been examined before the A–E program was designed, so this entire layer is classified as retrospective decision-value evidence. Within that retrospective program, however, each stage's question, thresholds, and decision rule were fixed before the stage was run. These studies determine which product functions remain defensible; they do not constitute a second independent OOS evaluation or strategy backtest.

How to read PASS and FAIL below. A label summarizes all required criteria fixed before that stage was run; it is not shorthand for a single p-value. A use fails when any required criterion is not met, even if some descriptive statistics move in the favorable direction.



6.2 Stage A: predictive downside signal — FAIL

The primary signal was B1's 20-day 5% Expected Shortfall (ES05) tail-loss severity for the eight-asset equal-weight portfolio. At each origin, ES05 was the mean cumulative simple return among the worst 5% of simulated 20-day portfolio outcomes, and risk severity was defined as $-$ES05, so a larger value meant worse simulated tail loss. This score was mapped into an expanding historical percentile initialized with the 2015–2018 validation origins and then updated only with strictly earlier eligible OOS origins. The high-risk state was fixed as the top quintile (Q5, percentile ≥ 80); the realized target was maximum-drawdown (MDD) severity over the next 20 trading days.

The rank association was directionally favorable but uncertain:

$$\rho_{\text{Spearman}} = 0.1676, \qquad \text{95\% confidence interval (CI)} = [-0.0330, 0.3440].$$

The Q5 minus Q1–Q4 MDD separation was also favorable at 0.38 percentage points, but its interval included zero. The additional required ordering against the lowest-risk quintile (Q1) failed as well: mean realized MDD severity was 2.80% in Q5 versus 3.91% in Q1. Thus the modest positive overall tendency was neither sufficiently robust nor consistently ordered to satisfy the qualification criteria. Stage A failed, so B1 ES05 is not used as a predictive downside warning.

6.3 Stage B-P: practical downside-control policy — FAIL

After Stage A had failed, the policy rule, thresholds, implementation lag, and cost assumption were fixed before a one-time execution on the 2019–2025 period whose primary OOS results were already known. A favorable result could have supported a bounded practical policy use, but it could not have upgraded the failed Stage-A predictive claim. The rule reduced risky exposure from 100% to 75% in the high-risk state and placed the remaining 25% in BIL, a short-term U.S. Treasury-bill ETF used only as a defensive cash-like sleeve. Execution was delayed by one full trading day. A 10-basis-point one-way cost was applied only to changes between the equal-weight risky sleeve and BIL; maintenance of that underlying sleeve was treated consistently across comparators.

Relative to remaining 100% in the eight-asset equal-weight portfolio, the dynamic policy reduced maximum drawdown from 22.85% to 20.54% and empirical daily ES05 loss severity from 1.72% to 1.51%, but compound annual growth rate (CAGR) fell from 10.03% to 8.61%. The dynamic policy averaged 91.31% risky exposure, so the exposure-matched static comparator held 91.31% in the equal-weight risky sleeve and 8.69% in BIL throughout. That static portfolio had MDD of 21.00% and CAGR of 9.41%. The dynamic

rule improved MDD by only about 0.46 percentage points relative to the static comparator, below the pre-specified requirement of about 1.05 percentage points, and it also delivered lower CAGR. The return-cost and timing-value conditions therefore failed.

> Reducing average risky exposure will mechanically reduce some downside. The exposure-matched static comparator asks whether *timing* adds enough value beyond that mechanical effect. It did not, so no dynamic BIL overlay is retained in the operational product.

6.4 Stage C: tail/risk-driver attribution — FAIL

Stage C asked whether the ex-ante scenario tail could identify the assets that would dominate a subsequently realized portfolio drawdown. It reused the 56 large-drawdown origins fixed by Stage A, defined using the 2015–2018 validation-period 90th-percentile MDD threshold (about 3.70%). For each simulated tail path, asset losses were measured over that path's own portfolio peak-to-trough interval; realized driver severities were measured over the realized portfolio drawdown interval. Three metrics compared predicted and realized drivers: E1, rank agreement across the eight assets; E2, recall of the two largest realized drivers; and E3, the fraction of realized positive asset-side loss captured by the predicted top two.

B1 and Flow were internally stable and both improved materially on a cheap trailing-60-day downside-ranking heuristic. The harder comparator was B0 historical block resampling, which was better on all three external-relevance metrics: E1/E2/E3 were 0.740/0.545/0.366 for B0, 0.708/0.473/0.342 for B1, and 0.687/0.438/0.334 for Flow F1. B1 and Flow also both missed the pre-specified E3 absolute-relevance threshold of 0.40. Stage C is therefore a **complexity test failure, not a stability failure**: B1's state-conditioned tail and Flow's neural tail did not justify their additional complexity over historical block resampling for forward driver attribution. Descriptive asset behavior may still be shown, but the system makes no claim of superior future loss-driver identification.

6.5 Stage D: scenario archetype / Stress Lab — PASS

Stage D changed the question. Instead of asking *when* risk would rise or *which asset* would drive a future loss, it asked whether the adverse scenario cloud contained a stable and externally relevant representation of the **joint stress pattern that actually occurred**. This is a narrower use of scenario generation.

At each origin, the stress tail consists of the worst 5% of 5,000 scenarios by equal-weight portfolio MDD: exactly 250 paths. Each path is classified from the eight assets' cumulative returns over that path's own portfolio peak-to-trough interval. Six mutually exclusive rule-based archetypes were fixed before the one-time Stage-D execution: broad cross-asset selloff; credit/REIT-led stress; commodity-up/duration-down; equity-bond joint selloff; equity selloff with bond hedge; and other tail. The subsequently realized 20-day stress path was classified by the same rules. These labels describe cross-asset return geometry, not causal macro regimes. The residual "other tail" category is retained deliberately so taxonomy inadequacy remains visible. Family shares are generated-scenario frequencies, not calibrated event probabilities.

Three external-relevance metrics were evaluated on the 56 fixed stress origins. D1 is the scenario frequency assigned to the family that actually occurred. D2 records whether that realized family was among the model's two largest tail families. D3 measures the standardized root-mean-square (RMS) distance between the realized multi-asset stress vector and the model's representative scenario from the same family; that representative is the actual simulated path closest to the family's mean standardized geometry. Simulation stability was assessed with five deterministic split-half checks of the same 5,000 paths. The reported total-variation distance (TVD) compares the resulting family-share vectors; lower TVD means less sensitivity to Monte Carlo randomness.

Table 3: Stage-D stress-archetype evidence on 56 fixed stress origins

Method	D1 share	D2 top-2	D3 distance	Stability TVD	Stage-D status
B0 block bootstrap	0.3661	0.9821	2.9171	0.0537	comparator
B1 EWMA-t	**0.6106**	0.9821	**1.8847**	**0.0468**	**pass**
Flow F1	0.3420	0.9821	1.9148	0.0498	does not qualify

Higher D1/D2 and lower D3/TVD are preferred. TVD is a split-half Monte Carlo stability diagnostic, not an external-relevance score.

Relative to B0, B1 increased D1 by 0.2445, left D2 unchanged, and reduced D3 by 35.4%. It therefore met the pre-specified rule requiring at least two of the three incremental margins to improve without a material reversal. It also cleared the separate safeguards for Monte Carlo stability, taxonomy adequacy, minimum standalone relevance, and leave-one-year-out robustness. Flow F1 was stable and met the standalone relevance safeguards, but versus B0 it reduced D3 by 34.4%, lowered D1 by about 0.024, and left D2 unchanged. Only one incremental margin improved, so Flow did not qualify; it also showed no incremental value over B1.

B1 therefore **passed the incremental stress-representation test against B0**; this is not a generative-model result. On the fixed 56-origin retrospective evaluation, B1 placed more tail-scenario mass on the stress archetype that subsequently occurred and produced a representative cross-asset stress geometry closer to the realized stress path than historical block resampling. The result supports B1 over B0 for this narrowly defined representation task, but it does not show incremental value from the modern generative challenger. A separate sparse-history novelty diagnostic was inconclusive because only one stress origin satisfied the pre-specified sparsity condition; Stage D therefore does not establish that B1 can reliably represent genuinely rare or previously unseen stress types.

6.6 Stage E: B1–Flow disagreement warning — FAIL

B1 and Flow often assigned materially different shares to the six stress archetypes. Their mean total-variation distance was 0.384. At the median origin, the cross-model disagreement was 5.86 times the larger of the two models' own Monte Carlo variability, so the difference between B1 and Flow was usually much larger than simulation noise within either model. Stage E asked a stronger practical question: *does a large B1–Flow disagreement warn that B1 is less reliable?*

The evidence answered no. If disagreement were a useful reliability warning, larger disagreement should be associated with *higher* subsequent B1 error or miss rates. Instead, on the 56 stress origins, larger disagreement was associated with *lower* B1 representative-geometry error (the Stage-D D3 distance; $\rho = -0.421$) and a lower realized-family miss measure, defined as one minus B1's D1 share assigned to the family that actually occurred ($\rho = -0.562$). The latter association had a 12-origin moving-block-bootstrap 95% interval of $[-0.834, -0.189]$, entirely in the adverse direction for the warning hypothesis. Across all 348 origins, the association with B1 CRPS was only 0.061, below the pre-specified support threshold. Disagreement also failed to add warning value beyond B1-only ambiguity measures, and the result was not robust to leaving out individual years. Stage E therefore fails as a reliability-warning use case. The system does not expose B1–Flow disagreement as a confidence badge and does not authorize a model switch, blend, or production threshold based on it.

Classification note. The initial automated report returned "partial structural only" because its classification routine checked the structural-disagreement condition before applying the pre-specified adverse-reversal rule. No numerical result was rerun. Applying the decision rule as written changes only the reviewed classification, to FAIL; the metrics, samples, bootstrap draws, and model outputs are unchanged.

7 Interpreting the Surviving Stage-D Result

7.1 Stress representation, not event prediction

The Stage-D result is useful precisely because its claim is narrower than predicting whether a crisis will occur. For each current market snapshot, B1 answers a conditional simulation question: given the return history available through that snapshot and the fixed EWMA-t specification, what adverse joint patterns appear in the simulated tail, and how do they differ from transparent historical resampling?

The output is therefore a structured stress map, not an event forecast. For example, if broad cross-asset selloff scenarios occupy a large share of the B1 tail while B0 is dominated by equity selloffs in which Treasuries hedge the loss, the contrast describes how the two scenario engines represent stress. It does not mean that the broad selloff has the displayed real-world probability of occurring.

7.2 Representative geometry rather than a synthetic average path

For every archetype with adequate tail representation, the system converts each asset's peak-to-trough cumulative log return into a standardized stress coordinate by dividing by the asset's trailing 60-day daily volatility times the square root of the interval length. It then selects the *actual simulated path* closest to that archetype's mean standardized stress vector. The representative is therefore a feasible joint scenario drawn from the simulation cloud, rather than a synthetic average path that may never have occurred in any simulated draw. The 60-day window is used only to normalize stress geometry; it is not B1's estimation window.

This design makes the stress narrative inspectable. A practitioner can see whether a broad selloff means "equities and credit fall while Treasuries also fail," whether commodities remain positive, or whether the historical comparator shows a materially different hedge pattern.

7.3 Why B0 remains visible

A conditional model should not replace the historical reference. B0 provides a different but valuable reference: stress patterns obtained by resampling actual contiguous 20-day history without conditioning on the current state. Showing B0 beside B1 provides a transparent anchor for interpretation. It reveals whether the conditional B1 tail resembles patterns already present in the historical record or assigns materially different weight to other joint stress geometries. The dashboard therefore presents the B1–B0 contrast rather than treating B1 as a standalone truth.

8 Operational Research System

8.1 On-demand workflow

The Stress Lab is designed for on-demand use rather than mandatory daily execution. A typical cycle is:

1. ingest the latest eligible operational market returns;
2. run data-quality checks and save a dated input snapshot together with hashes, which act as digital fingerprints for later integrity checks;
3. fix the eligible return history as of that snapshot and separately compute the trailing-60 volatility scales used only by the stress-geometry diagnostics;
4. generate B1 20-day scenarios from the eligible history under the fixed EWMA-t specification;
5. compute portfolio tail/path diagnostics and the Stage-D stress-archetype outputs;
6. generate the B0 20-day historical block-bootstrap comparator from the same eligible history;
7. apply each configured portfolio's weights to the same B1 scenario cloud;
8. save the scenario bundle, analytics outputs, and audit/validation records;
9. refresh the dashboard.

Flow may still be computed in the research-only monitoring workflow, but after Stage E no operational decision, warning, or practitioner-facing claim depends on it. The core Stress Lab interpretation is B1 plus the B0 historical comparator. Operator instructions are maintained in the Full Manual [8].

The refresh is fail-closed: a failed data update, data-quality check, or scenario issuance cannot appear as a successful current run. A previously saved snapshot can still be opened explicitly in dashboard-only mode, but its original data-as-of and issue timestamps remain visible and it must not be represented as a newly refreshed result.

8.2 Practitioner interface

Figure 1 shows the primary practitioner-facing Stress Lab screen. The top row reports the dominant archetype within the current B1 stress tail, its generated-scenario share, the corresponding B0 historical-resampling share, and the portfolio's model-implied 5% expected shortfall (ES). Here, "current" means conditional on the data snapshot used for that run; it does *not* mean that the displayed archetype is predicted to occur. Scenario shares are generated frequencies, not calibrated event probabilities, and the B1-minus-B0 share gap is a descriptive contrast in modeled stress structure rather than a directional market signal.

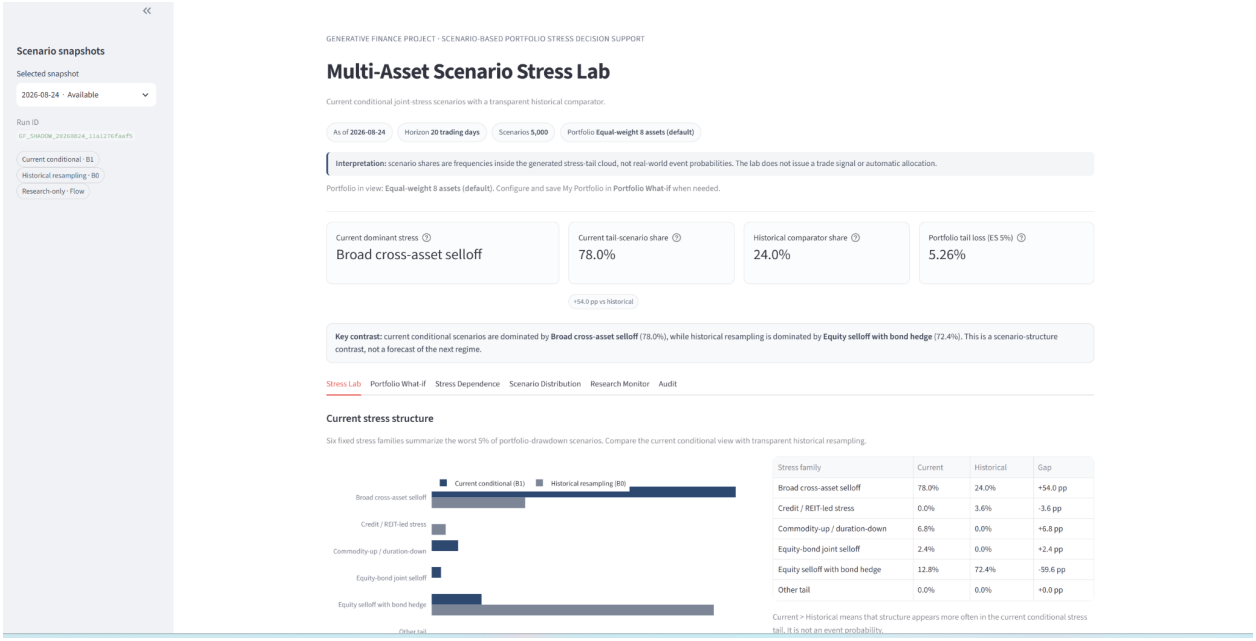


Figure 1: Primary Stress Lab dashboard surface. The current B1 conditional stress structure is shown beside the B0 historical-resampling comparator. The interface avoids trade-signal and calibrated-probability language.

8.3 Portfolio what-if analysis

A baseline portfolio and a candidate alternative are evaluated on exactly the *same* 5,000 B1 paths; changing the weights does not regenerate the market scenario set. This isolates portfolio-weight effects from scenario-generation effects. The interface reports the median 20-day return, 5% return quantile, 5% ES, 5% drawdown quantile, and stress-family representative-path outcomes for both portfolios.

The feature is descriptive and human-in-the-loop. It does not optimize weights, recommend a portfolio, or search backward for weights that minimize displayed scenario loss. A candidate that looks better under the fixed scenario cloud is therefore a what-if result, not evidence of future realized outperformance.

8.4 Reproducibility and audit trail

Each official run records its run identifier; dated input snapshot and hashes; data-source boundary; model identifiers, roles, and material parameters; issue time and data-as-of; scenario count and horizon; deterministic seed policy; output-artifact hashes; and model health/latency information. Flow checkpoint metadata are also recorded when Flow is computed for research monitoring. Together, these fields provide an audit identity showing which data, model configuration, randomization policy, and artifacts produced the issued analysis.

The record supports verification and reconstruction, but not a guarantee of exact byte-for-byte regeneration across platforms. Exact regeneration also depends on retaining compatible source code, software libraries, the execution environment, and any required checkpoints. Portfolio weights are stored separately: a saved My Portfolio preserves only the latest saved weights and update timestamp, while an unsaved candidate is not versioned. Reconstructing an exact historical what-if screen therefore also requires the weights used at that time to have been preserved.

9 Claim Boundaries and Public Product Scope

9.1 Fail-closed scope rule

In this project, *fail-closed* means that when a pre-specified use-case criterion fails, the proposed feature is removed or narrowed rather than being re-engineered on the same evidence until something passes. After the A–E program was completed, the 2019–2025 interval was not reused to search for a different risk threshold, BIL intensity, tail-driver definition, archetype taxonomy, disagreement measure, lookback, horizon, or new generator merely to rescue a failed use case.

This rule matters because an interactive dashboard creates many opportunities to choose features after seeing the results. With enough thresholds, labels, and charts, a persuasive pattern can almost always be found after the fact. The public product therefore reflects the analyses that were actually specified and evaluated, including failed features that were excluded.

9.2 Allowed and prohibited claims

Table 4: Public claim boundary

Allowed	Not allowed
Generated scenario-family shares within the B1 modeled stress tail	Calibrated probabilities that a crisis or archetype will occur
B1 conditional stress structure compared with B0 historical resampling	Market-timing or crash prediction inferred from the share gap
Representative joint stress paths and geometry	Causal macro or regime labels inferred from an archetype name
Model-implied VaR/ES/MDD and same-cloud portfolio diagnostics	Treating those diagnostics as validated forecasts of realized loss
B0 as a transparent historical comparator	Claim that B1 is universally superior to historical simulation
Flow F1 as a research-only challenger	Claim that a deep generative model outperforms B1
Human portfolio what-if comparison under a fixed scenario cloud	Predictive traffic-light signals, automated BIL overlays, or model-switch/blend rules

9.3 Why failed uses remain documented

The failed Stage A, B-P, C, and E evaluations are part of the evidence that defines the product scope. Each excluded capability is traceable from a proposed use to a pre-specified test, a failed criterion, and its removal from the public operational feature set. Preserving that chain makes the design consequences of negative evidence auditable and prevents unsupported prototypes from being retained on the dashboard.

10 Limitations and Research Boundaries

The current system has several material limitations.

First, the Stage-D evidence is retrospective. It uses 56 fixed large-drawdown origins from the same 2019–2025 period that had already been used for the primary OOS evaluation. The Stage-D taxonomy and decision criteria were fixed before Stage D was run, but the evaluation period itself was not fresh. The evidence therefore supports only the narrow claim that B1 outperformed B0 on this retrospective stress-archetype representation test and justified the bounded Stress-Lab use retained in the product. It is not independent prospective validation.

Second, the taxonomy is deliberately simple. Six rule-based archetypes favor interpretability and ease of audit over a richer representation. A learned clustering model might identify more nuanced states, but searching alternative taxonomies on the same 2019–2025 evidence would make the taxonomy a post-hoc choice and invalidate the current Stage-D interpretation.

Third, the universe is deliberately compact. The eight ETFs cover major exposures across equity, duration, credit, commodities, gold, and real estate, but not every position in a broad institutional allocation. High yield, long-duration government bonds, currencies, alternatives, and derivatives can change both the magnitude and the sign of cross-asset stress dependence. The reported results therefore should not be assumed to transfer unchanged to broader portfolios.

Fourth, B1 is intentionally simple. At each forecast origin, B1 estimates a mean-return vector and EWMA covariance matrix from the eligible history, then holds both fixed throughout each simulated 20-day path while drawing heavy-tailed multivariate Student-t shocks day by day. This captures the current covariance environment and heavy tails, but not every nonlinear transition, asymmetry, intrahorizon volatility-regime change, or state-dependent dependence pattern. Its use in the Stress Lab is justified by comparative evidence, not by theoretical completeness.

Fifth, operational continuity uses a different data source. The research evidence uses the fixed CRSP return panel. Current runs use Yahoo adjusted prices after the documented 2025 overlap comparison. This bridge supports ongoing decision support, but the operational and research sources are not identical. Yahoo history should not be treated as a reconstruction of the CRSP research panel. The distinction remains part of the system boundary.

Sixth, portfolio what-if analysis is model-conditional. A candidate portfolio that looks better inside the current B1 scenario cloud is not guaranteed to outperform in realized returns. The tool is designed to expose scenario trade-offs under a common modeled market environment, not to certify optimal weights.

Future work should begin with a materially new hypothesis and, where the claim requires it, genuinely new evidence rather than reusing 2019–2025 to rescue a failed use case. Potential extensions include broader asset universes, new state representations, separately specified macro inputs, learned stress taxonomies, prospective monitoring, or scenario ensembles. None is part of the current public claim set.

11 Conclusion

The Multi-Asset Scenario Stress Lab was shaped by a sequence of evidence-driven reductions. Modern neural generators did not displace the conventional EWMA-t reference in primary OOS: Flow F1 remained close to B1 but did not establish superiority, while CVAE and DDPM were not retained for operational use. B1 tail severity did not qualify as a predictive downside warning. A dynamic defensive overlay improved raw downside metrics but failed the required return-cost and exposure-matched timing-value criteria. Conditional tail-driver analysis did not outperform B0 historical block resampling. B1–Flow disagreement was clearly larger than within-model simulation noise, but it did not behave as a reliability warning.

Only one practical use qualified: a transparent scenario-archetype Stress Lab. Under the six-family taxonomy and 56 stress origins fixed before the one-time Stage-D execution, B1 increased the mean scenario share assigned to the subsequently realized family by 0.2445 (24.45 percentage points) relative to B0 and reduced representative-geometry distance by 35.4%; D2 was unchanged. B1 also met the separate safeguards for simulation stability, taxonomy adequacy, minimum standalone relevance, and leave-one-year-out robustness. Within this explicitly retrospective Stage-D design, the evidence supports retaining a narrow human-in-the-loop Stress Lab. It does not support a broader trading, market-timing, or calibrated-probability engine.

The broader lesson is methodological rather than model-specific. Practical value in generative-finance research must be demonstrated for the intended decision use, against a credible simple comparator, with explicit information timing and claim boundaries that contract when evidence fails. Here, that discipline produced a narrower system than initially contemplated—and a more defensible one.

A Technical Definitions

A.1 Portfolio path, drawdown, and stress tail

For portfolio weights w_i and simple asset returns $r_{i,t}^{(s)}$, the portfolio path is generated from the same scenario cloud. Let $W_t^{(s)}$ be cumulative wealth. Drawdown and maximum drawdown are

$$DD_t^{(s)} = 1 - \frac{W_t^{(s)}}{\max_{u \le t} W_u^{(s)}}, \qquad MDD^{(s)} = \max_t DD_t^{(s)}.$$

The Stage-D stress tail consists of the 250 scenarios (5% of 5,000) with the largest maximum-drawdown severity under the equal-weight eight-asset portfolio used to define the stress event.

A.2 Representative stress geometry

For a scenario's portfolio peak-to-trough interval of length L, asset i receives the standardized stress value

$$Z_i = \frac{\sum_{t \in \text{P2T}} r^{\log}_{i,t}}{\sigma_{i,60}\sqrt{L}},$$

where P2T denotes the trading days in that scenario's equal-weight portfolio peak-to-trough interval and $\sigma_{i,60}$ is trailing 60-day daily log-return volatility at the origin. Within an archetype, the representative path is the actual simulated path whose Z vector is closest in RMS Euclidean distance to the archetype mean vector.

A.3 Stage-D qualification criteria

A candidate qualified on incremental value only if at least two of three margins improved relative to B0: D1 by at least +0.05, D2 by at least +0.05, or D3 by at least 10%. A material reversal was also prohibited: D1 and D2 differences could not fall below -0.05, and D3 could not be more than 10% worse than B0. Candidate retention additionally required four safeguards. Monte Carlo stability required mean split-half TVD no greater than 0.10 and mean top-two-family overlap of at least 0.80. Taxonomy adequacy required the realized "other-tail" fraction to be no greater than 0.35. Absolute relevance required mean D1 of at least 0.15, D2 of at least 0.60, D3 no greater than 2.00, and mean simulated "other-tail" share no greater than 0.40. Leave-one-year-out robustness prohibited any omitted-year result with D1 below 0.08, D2 below 0.45, or D3 above 3.00. B1 met the D1 (+0.2445) and D3 (-35.4%) incremental margins, D2 was unchanged, and all four safeguards passed.

B Model and Use-Case Outcomes

Object	Final role/status	Interpretation
B0 block bootstrap	Historical stress comparator	Transparent resampling reference; not the core conditional engine.
B1 EWMA-t	Reference model / Stress Lab core	Strongest overall primary-OOS calibration and dependence evidence; passed Stage D against B0 for the retained stress-archetype use.
CVAE	Not retained at validation	Validation performance on joint-distribution and dependence-sensitive metrics was not competitive with B1; retained only as a documented research reference.
DDPM	Not retained after primary OOS	Competitive at validation, but primary OOS showed no compensating advantage over B1.
Flow F1	Research-only challenger	Competitive with B1 on the main primary-OOS scores; lower mean Variogram score and favorable descriptive path metrics, but block-bootstrap inference did not establish superiority.
Flow F3	Higher-step research reference	40-step Flow specification; primary OOS showed no compensating quality benefit over the much faster 10-step F1 configuration.
Stage A	FAIL	No predictive downside-warning indicator.
Stage B-P	FAIL	No automated BIL de-risking overlay.
Stage C	FAIL	No validated claim of superior future tail-driver attribution.
Stage D	PASS	Passed only for stress-archetype representation against B0; narrow Stress Lab use retained.
Stage E	FAIL	No B1–Flow reliability warning, switching, or blending rule.

C References

[1] Gneiting, T., and Raftery, A. E. (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association*, 102(477), 359–378. doi: 10.1198/016214506000001437.

[2] Scheuerer, M., and Hamill, T. M. (2015). "Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities." *Monthly Weather Review*, 143(4), 1321–1334. doi: 10.1175/MWR-D-14-00269.1.

[3] Künsch, H. R. (1989). "The Jackknife and the Bootstrap for General Stationary Observations." *The Annals of Statistics*, 17(3), 1217–1241. doi: 10.1214/aos/1176347265.

[4] Engle, R. (2002). "Dynamic Conditional Correlation: A Simple Class of Multivariate Generalized Autoregressive Conditional Heteroskedasticity Models." *Journal of Business & Economic Statistics*, 20(3), 339–350. doi: 10.1198/073500102288618487.

[5] Sohn, K., Lee, H., and Yan, X. (2015). "Learning Structured Output Representation Using Deep Conditional Generative Models." *Advances in Neural Information Processing Systems* 28, 3483–3491.

[6] Ho, J., Jain, A., and Abbeel, P. (2020). "Denoising Diffusion Probabilistic Models." *Advances in Neural Information Processing Systems* 33.

[7] Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. (2023). "Flow Matching for Generative Modeling." *International Conference on Learning Representations (ICLR)*. arXiv: 2210.02747.

[8] Generative Finance Project (2026). *Multi-Asset Scenario Stress Lab Full Manual*, v1.1. Companion operational manual; validation boundaries are governed by the system's validation records.