DIACHRONIC SAMPLE INTEGRATION: ROBUST TAIL-RISK ESTIMATION WITH GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep generative models are increasingly used as simulators for downstream decision making under data scarcity, but in risk-sensitive applications their usefulness depends on rare adverse scenarios rather than typical samples. Standard generative objectives prioritize bulk distributional fidelity, leaving low-probability tails vulnerable to localized optimization noise and making tail-dependent functionals unstable under finite simulation budgets. We introduce Diachronic Sample Integration (DSI), a test-time inference framework that ensembles generated samples across checkpoints from a stochastic training trajectory. DSI targets a checkpoint-mixture distribution that averages checkpoint-specific tail fluctuations rather than relying on a single brittle endpoint. We formalize this mechanism through a finite-budget bias-variance theory. Empirically, across multivariate synthetic processes and high-frequency trading data, DSI substantially reduces tail-estimation error compared to single-checkpoint baselines under fixed simulation budgets, outperforming standard diffusion and state-of-the-art tail-aware baselines without modifying the generative objective.

1 INTRODUCTION

Deep generative models are increasingly used as simulators for downstream decision making under data scarcity. In risk-sensitive applications, however, simulator reliability depends not only on typical samples but also on rare adverse scenarios. Reliable estimation of statistical functionals defined on low-probability regions is central to risk-sensitive inference. Because empirical tail observations are exceptionally scarce, practitioners increasingly rely on deep generative models to act as approximate simulators for downstream risk estimation (Gonen et al., 2025; Tanaka et al., 2025; Li et al., 2025).

However, standard generative objectives, such as score matching or maximum likelihood, prioritize the "bulk" of the probability mass where data is dense (Pandey et al., 2025). The extreme tails are consequently starved of reliable gradient signals and become dominated by localized optimization noise. Drawing synthetic samples from a single "converged" checkpoint often yields severe tail miscalibration, either arbitrarily truncating the tail (mode collapse) or hallucinating unrealistic extremes. Existing methods pursue tail fidelity through risk-specific objectives or additional structural and distributional assumptions (Huang et al., 2024; Galib et al., 2024), which may encounter frictions and usage restrictions in regulated application environments, as reflected in guidance from the European Central Bank (2026, Market risk-CRR3, p. 293), the Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency (2011, pp. 3–4, 10), and the Basel Committee on Banking Supervision (2019).

We therefore study the same tail-estimation problem from a complementary inference perspective: finite-budget risk estimates without imposing any parametric tail form, distributional constraint, or additional training objective. In many risk-management settings, the immediate task is not to recover an asymptotic tail law, but to estimate risk measures such as Value-at-Risk (VaR) and Expected Shortfall (ES) under a limited simulation budget constrained by compute or validation resources. We therefore propose Diachronic Sample Integration (DSI)[1], a test-time framework that ensembles

[1]The source code for this project is available at https://anonymous.4open.science/r/DSI-Submission/.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

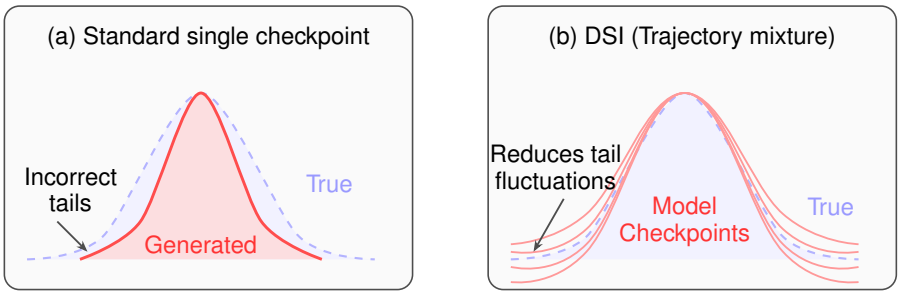


Figure 1: Conceptual illustration of DSI. (a) Single checkpoints prioritize bulk distributional fidelity, often truncating extreme tails due to optimization noise. (b) DSI aggregates heterogeneous checkpoints across a single trajectory, the resulting mixture averages checkpoint-specific tail fluctuations while any persistent trajectory bias remains as an irreducible floor.

samples across stochastic training checkpoints (Figure 1). Rather than treating a single converged model state as the sole simulator, DSI uses the training trajectory as a source of checkpoint-level variability and aggregates generated samples to reduce tail-estimation instability while leaving the underlying generator unchanged.

DSI is not guaranteed to remove systematic model bias. Its statistical benefit arises when post-burn-in checkpoint errors contain averageable fluctuations around a persistent trajectory bias. We therefore analyze DSI through a finite-budget bias-variance decomposition for VaR and ES, showing that checkpoint stride and mixture size enter through an effective checkpoint count determined by tail-error autocorrelation. Empirically, on controlled synthetic processes and high-frequency NASDAQ data, DSI substantially reduces relative tail-estimation error compared to standard single-checkpoint diffusion models under matched sampling budgets, outperforming TailGAN (Cont et al., 2025), the current state-of-the-art tail-aware generative benchmark.

This paper makes the following contributions:

- We introduce DSI, a simple lightweight test-time sample-integration framework that improves tail-risk inference via checkpoint ensembling.
- We derive a bias-variance theory that characterizes when trajectory mixing improves tail estimation and when persistent bias or checkpoint dependence limits its benefit.
- We show that DSI substantially reduces tail-estimation error and estimator variance under fixed simulation budgets, without retraining or changing the generative objective.

2 Related Work

2.1 Generative Simulators and Tail-Sensitive Objectives

Deep generative models are increasingly employed as probabilistic simulators to mitigate data scarcity in risk-sensitive inference. Standard architectures (Takahashi & Mizuno, 2025; Yuan & Qiao, 2024; Xia et al., 2024; Yoon et al., 2019) optimize global distributional fidelity, which can leave low-probability regions vulnerable to localized optimization noise (Liu et al., 2024; Kollovieh et al., 2023).

Recent efforts to improve synthetic tail calibration have introduced adversarial risk penalties, extreme-aware diffusion mechanisms, or parametric extreme-value constraints (Cont et al., 2025; Galib et al., 2024; Huang et al., 2024). Recent theory further shows that heavy-tailed target distributions pose distinct statistical challenges for score-based diffusion models: under polynomial tail decay, score-estimation and sampling rates explicitly depend on the tail index, unlike the exponentially decaying regime (Yu & Yu, 2026). A related but different line studies class-imbalanced long-tailed generation, where tail labels are underrepresented. For example, Das et al. (2025) propose a training-time mutual-learning regularizer for conditional diffusion models to improve worst-case class-label performance.

These approaches address tail behavior by modifying the training objective, the modeling assumptions, or the conditional learning problem. In contrast, DSI preserves the original generative objective and addresses checkpoint-level tail instability at inference time via sample-space aggregation and trajectory-level uncertainty averaging.

2.2 Model Ensembling and Averaging

Ensembling methods such as Deep Ensembles (Lakshminarayanan et al., 2017), Stochastic Weight Averaging (SWA) (Izmailov et al., 2018), and snapshot ensembles (Huang et al., 2017) aggregate parameters or predictions to improve predictive performance and generalization. These approaches operate in parameter or predictor space by averaging network weights (which can destroy complex learned manifolds) or point-predictions.

While some of these methods leverage multiple independent runs or training trajectories to find flatter optima (Wortsman et al., 2022; Maddox et al., 2019; Garipov et al., 2018), DSI shifts the focus to the sample space. By pooling stochastic samples across checkpoints, DSI constructs a macroscopic mixture distribution to stabilize non-linear statistical functionals. To address the high correlation often found between consecutive iterates in a single trajectory, we treat checkpoint spacing as a design variable via a temporal stride. This mechanism is designed to increase the effective checkpoint count by reducing autocorrelation among tail errors, a structural hypothesis we validate empirically in Section 5.

3 Problem Formulation

3.1 Probability Space and Tail-Dependent Functionals

We now formalize the downstream inference problem induced by a learned generative simulator: the simulator produces synthetic paths, which are mapped to a scalar decision quantity whose lower tail determines risk.

To formally ground the problem of tail estimation, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X : \Omega \to \mathbb{R}$ be a random variable representing a scalar quantity of interest. In risk-sensitive applications, X typically models the future return of a financial portfolio or the profit-and-loss (PnL) of a trading strategy, where large negative realizations correspond to severe and consequential losses. We denote the true, unknown data-generating distribution of X by $P^\star$.

We are interested in estimating a statistical functional $\psi : \mathcal{P} \to \mathbb{R}$. Conceptually, ψ is a mapping that takes an entire probability distribution as its input and outputs a single real-valued scalar (such as a risk metric). Here, $\mathcal{P}$ explicitly denotes the *domain* of this functional, the class of admissible probability distributions over $\mathbb{R}$ for which ψ is mathematically well-defined. For instance, evaluating Expected Shortfall strictly requires $\mathcal{P}$ to be restricted to the family of distributions possessing a finite first moment. We assume throughout that both the true target distribution $P^\star$ and our learned generative approximations belong to $\mathcal{P}$.

Throughout this work, we use Value-at-Risk (VaR) and Expected Shortfall (ES) as canonical examples of such functionals (Artzner et al., 1999; Rockafellar & Uryasev, 2000).

For a given tail probability level $\alpha \in (0, 1)$ (typically small, e.g., $\alpha = 0.01$ or 0.05), the α-quantile, or Value-at-Risk, of a distribution P with cumulative distribution function (CDF) F_P is defined as:

$$q_\alpha(P) := \inf\{x \in \mathbb{R} : F_P(x) \geq \alpha\}. \tag{1}$$

We use $\mathrm{VaR}_\alpha(P)$ and $q_\alpha(P)$ interchangeably for this lower-tail quantile.

Intuition for VaR: In risk management, Value-at-Risk (VaR) provides a severity threshold at a specified confidence level. For example, a 5% lower-tail VaR identifies a loss threshold exceeded in approximately the worst 5% of scenarios. VaR has historically been central to the Basel international banking regulatory framework, which sets global standards for bank capital and market-risk measurement Basel Committee on Banking Supervision (2006). However, VaR is merely a boundary: it is blind to the shape of the distribution beyond that threshold. It does not quantify how much the distribution can deteriorate in the extreme tail, nor the magnitude of losses once the VaR threshold is breached.

To capture the actual severity of extreme events, we use Expected Shortfall (ES) at level α in its lower-tail CDF form:

$$\mathrm{ES}_\alpha(P) := q_\alpha(P) - \frac{1}{\alpha}\int_{-\infty}^{q_\alpha(P)} F_P(x)\,dx, \tag{2}$$

whenever the integral is finite. If $Y \sim P$ and P is continuous with $F_P(q_\alpha(P)) = \alpha$, this equals $\mathbb{E}_P[Y \mid Y \le q_\alpha(P)]$; for empirical distributions we use the corresponding fractional-order-statistic version.

Intuition for ES: Expected Shortfall (ES) measures the average severity of outcomes in the extreme α-tail. While VaR tells us the threshold of a "bad day," ES tells us the average severity of losses on those bad days. This distinction is operationally important: the Basel III/FRTB market-risk framework Basel Committee on Banking Supervision (2019) moved from VaR toward ES as the headline internal-model risk measure because ES captures tail severity beyond the VaR cutoff. Because ES depends on the entire tail rather than a single point, it provides a more comprehensive measure of risk, but is significantly more difficult to estimate reliably from limited empirical data.

In many modern settings, direct empirical estimation of $\psi(P^\star)$ is infeasible due to the sheer scarcity of tail observations. Instead, practitioners increasingly rely on synthetic samples generated by a learned model P_θ to support downstream risk inference.

3.2 Generative Approximation and the Single-Checkpoint Bottleneck

Let Θ denote the parameter space of the generative simulator, and let $\mathcal{M} = \{P_\theta : \theta \in \Theta\} \subseteq \mathcal{P}$ denote the corresponding parametric class of generative distributions. We assume a mapping $\theta \mapsto P_\theta$ and fit the simulator by minimizing a global objective $\mathcal{L}(\theta; \mathcal{D})$ over a training dataset $\mathcal{D}$. Stochastic optimization traverses a complex loss landscape, inducing a training trajectory of model parameters $\theta_1, \theta_2, \ldots, \theta_T$, where each iterate θ_t defines a specific generative distribution $P_t := P_{\theta_t}$.

Standard practice selects a single "converged" checkpoint (typically θ_T) and draws a synthetic dataset $X_1, \ldots, X_N \sim P_T$. This yields the plug-in estimator $\hat{\psi}_T := \psi(\hat{P}_T)$, where $\hat{P}_T$ is the empirical distribution of the generated samples.

However, relying on a single checkpoint is structurally brittle for tail estimation. Because global generative objectives (such as maximum likelihood or score matching) prioritize the "bulk" of the probability mass where data is dense, the tail regions of P_t receive very little reliable gradient signal. Consequently, the tail behavior is highly sensitive to the optimization noise of the final training steps. A single checkpoint might arbitrarily truncate the tail (mode collapse) or hallucinate unrealistic extremes, leading to highly unreliable VaR and ES estimates.

3.3 Checkpoint Ensembling on Generated Samples

To overcome the instability of single-checkpoint estimation, we treat the stochastic training trajectory not as a path to a single point estimate, but as a rich ensemble of tail approximations. Our core premise is that checkpoint ensembling on the generated samples leads to substantially better and more stable tail estimation.

Rather than discarding intermediate checkpoints, we aggregate samples drawn across them. Let $S \subset \{1, \ldots, T\}$ be a selected set of checkpoints, and let n_t denote the number of samples drawn from each checkpoint P_t, subject to a fixed total sampling budget $\sum_{t \in S} n_t = N$. Defining weights $w_t := n_t/N$, pooling these generated samples naturally targets the trajectory-induced mixture distribution. For the theoretical analysis and experiments we focus on the equal-allocation case $w_t = 1/K_S$, corresponding to Algorithm 1 in Appendix A.1, but the framework is general and can accommodate any allocation scheme. The resulting mixture distribution is

$$\bar{P}_S = \sum_{t \in S} w_t P_t. \tag{3}$$

The Diachronic Sample Integration (DSI) estimator is computed directly on the ensembled samples:

$$\hat{\psi}_{\mathrm{DSI}} := \psi(\hat{P}_S), \tag{4}$$

where $\hat{P}_S$ is the empirical distribution of the pooled samples.

Why Ensembling Improves Tail Estimation: By ensembling generated samples across the temporal trajectory, DSI averages checkpoint-specific tail errors that fluctuate across training. An isolated tail failure at one checkpoint can be diluted by other checkpoints, provided those failures are not perfectly correlated and do not share the same systematic direction. The mixture distribution $\bar{P}_S$ therefore reduces fluctuation-driven estimator variance while leaving persistent trajectory bias as an irreducible floor, all without modifying the underlying generative architecture or requiring additional training runs.

3.4 Checkpoint Subsampling and Stride

Consecutive checkpoints along a stochastic training trajectory often induce highly correlated distributions. Pooling many nearby checkpoints may therefore introduce redundancy without significantly diversifying the simulated tail events.

To control correlation and maximize the informational value of the mixture, we parameterize S using a stride $M \geq 1$ and a checkpoint count K:

$$S(M, K) = \{t_0,\ t_0 + M,\ t_0 + 2M,\ \ldots,\ t_0 + (K-1)M\},$$

where t_0 is chosen after the model has completed its initial burn-in phase, as detailed in Appendix A.2.

If the trajectory evolves smoothly under a distributional metric, consecutive checkpoints induce highly correlated components. Increasing the stride M increases the separation in parameter space between retained checkpoints and can increase the effective checkpoint count by lowering tail-error autocorrelation. As K grows and the mixture spans a larger portion of the trajectory, the marginal impact of the stride naturally diminishes. We study this effect empirically in Section 5.

4 Theoretical Analysis

We now give a finite-budget bias-variance account of DSI. DSI targets the checkpoint-mixture distribution $\bar{P}_S$, not $P^\star$ directly. Its usefulness therefore rests on two conditions: the mixture $\bar{P}_S$ must remain close to the true target distribution $P^\star$, and checkpoint-specific tail errors must fluctuate sufficiently along the training trajectory. Under these conditions, pooling samples across spaced checkpoints reduces the fluctuation-driven component of tail-functional estimation error, while any persistent mixture bias relative to $P^\star$ remains as an irreducible floor.

Setup. Let $F^\star$ denote the CDF of the true scalar target distribution $P^\star$, and let F_t denote the CDF induced by checkpoint t. For a selected checkpoint set S, DSI targets the checkpoint-mixture CDF $\bar{F}_S(x) := |S|^{-1} \sum_{t \in S} F_t(x)$, equivalently the mixture distribution $\bar{P}_S := |S|^{-1} \sum_{t \in S} P_t$. The empirical DSI estimator draws a fixed total budget of synthetic samples across checkpoints in S and applies the plug-in tail functional to the pooled empirical distribution $\widehat{P}_{S,N}$.

With $q^\star := q_\alpha(P^\star)$ denoting the true lower-tail quantile, the VaR and ES checkpoint-level tail-error variables are

$$U_t^q = F_t(q^\star) - F^\star(q^\star), \tag{5}$$

$$U_t^{ES} = \int_{-\infty}^{q^\star} \big(F_t(x) - F^\star(x)\big)\, dx. \tag{6}$$

Under the post-burn-in checkpoint-error assumption in Appendix B.1, each target functional $\psi \in \{q_\alpha, \mathrm{ES}_\alpha\}$ has an associated checkpoint-error process $U_t^\psi = b_\psi + \xi_t^\psi$, where b_ψ is the persistent bias and ξ_t^ψ is a mean-zero weakly stationary fluctuation with variance σ_ψ^2 and autocorrelation $\rho_\psi(h)$. The effective checkpoint count for ψ, with $K_S = |S|$, is

$$K_{\mathrm{eff},\psi}(K_S, M) = \frac{K_S}{1 + 2 \displaystyle\sum_{\ell=1}^{K_S - 1} \left(1 - \frac{\ell}{K_S}\right) \rho_\psi(\ell M)}. \tag{7}$$

Thus $K_{\mathrm{eff},\psi}$ is the number of effectively independent checkpoint errors for the functional being estimated: it is close to K_S when spaced checkpoint errors are weakly correlated and close to one when they are highly correlated. In the theorem below we apply this construction to $\psi = q_\alpha$ and $\psi = \mathrm{ES}_\alpha$, writing $K_{\mathrm{eff},q}$ and $K_{\mathrm{eff},ES}$ and suppressing their (K_S, M) arguments. The empirical diagnostics in Appendix D.2 estimate these quantities for the trained trajectories.

The assumptions have three roles. Assumption 1 gives local tail regularity, ensuring that small CDF perturbations near $q^\star$ translate smoothly into VaR and ES perturbations. Assumptions 2 and 4 describe the post-burn-in checkpoint-error process and the local regularity of the selected checkpoint mixture, separating persistent trajectory bias from averageable checkpoint fluctuations. Assumptions 3 and 5 handle the finite-budget stratified sampling layer: DSI samples are independent conditional on checkpoints but not identically distributed marginally, and they estimate the mixture $\bar{P}_S$, not a single checkpoint. Full technical statements are given in Appendix B.1.

Theorem 1 (Finite-budget DSI MSE expansion) *Let* $\widehat{q}^{DSI}_{\alpha,S,N} = q_\alpha(\widehat{P}_{S,N})$ *and* $\widehat{\mathrm{ES}}^{DSI}_{\alpha,S,N} = \mathrm{ES}_\alpha(\widehat{P}_{S,N})$. *Under the assumptions stated in Appendix B.1,*

$$\mathbb{E}\Big[\big(\widehat{q}^{DSI}_{\alpha,S,N} - q^\star\big)^2\Big] = \frac{1}{f^\star(q^\star)^2}\left(b_q^2 + \frac{\sigma_q^2}{K_{\mathrm{eff},q}}\right) + L^{sam}_{q,N} + o(\cdot), \tag{8}$$

$$\mathbb{E}\Big[\big(\widehat{\mathrm{ES}}^{DSI}_{\alpha,S,N} - \mathrm{ES}_\alpha(P^\star)\big)^2\Big] = \frac{1}{\alpha^2}\left(b_{ES}^2 + \frac{\sigma_{ES}^2}{K_{\mathrm{eff},ES}}\right) + L^{sam}_{ES,N} + o(\cdot). \tag{9}$$

Here $L^{sam}_{q,N}$ *and* $L^{sam}_{ES,N}$ *are the stratified finite-sample plug-in variances for the fixed checkpoint mixture. Their exact forms and the full two-stage theorem statement are given in Appendix B.2, with the proof in Appendix B.3.*

Interpretation. The expansion separates DSI error into a trajectory term and a sampling term. For a tail functional ψ, the trajectory term $b_\psi^2 + \sigma_\psi^2/K_{\mathrm{eff},\psi}$ decomposes into a persistent bias floor b_ψ^2 and a checkpoint-fluctuation component $\sigma_\psi^2/K_{\mathrm{eff},\psi}$ that decreases as $K_{\mathrm{eff},\psi}$ grows. DSI targets the checkpoint mixture, not $P^\star$ directly, so bias is not removed by averaging. Stride M helps only by reducing autocorrelation, increasing K helps only while it increases the functional-specific $K_{\mathrm{eff},\psi}$, and very large K is eventually limited by floor allocation and the constraint $K_S \le N$.

Two immediate consequences make the theorem operational: it predicts when DSI should help and when trajectory averaging must fail.

Corollary 1 (Mechanistic regimes). Theorem 1 identifies three regimes for each tail functional ψ: (i) a variance-reducible regime where b_ψ^2 is small and $K_{\mathrm{eff},\psi}$ grows, (ii) a correlation-limited regime where checkpoint fluctuations are present but highly autocorrelated, so $K_{\mathrm{eff},\psi} \ll K_S$, and (iii) a bias-limited regime where b_ψ^2 dominates and DSI cannot remove the leading error term. The proof is given in Appendix B.4.

Proposition 1 (Failure modes). DSI cannot correct perfectly correlated checkpoint errors, shared directional lower-tail misspecification across all selected checkpoints, or persistent bias floors. In particular, if $\rho_\psi(\ell M) = 1$ for all $\ell = 1, \dots, K_S - 1$, then $K_{\mathrm{eff},\psi} = 1$. If, for some sign $s \in \{-1, 1\}$ and nonnegative function d, $s\{F_t(x) - F^\star(x)\} \ge d(x)$ for all $t \in S$ on a tail interval, then $s\{\bar{F}_S(x) - F^\star(x)\} \ge d(x)$ on that interval. Even when $K_{\mathrm{eff},\psi} \to \infty$ and $N \to \infty$, the first-order VaR and ES MSE floors remain $b_q^2/f^\star(q^\star)^2$ and b_{ES}^2/α^2, respectively. The proof is given in Appendix B.5.

These regimes motivate the diagnostics in Appendix D.2, including Figure 6, and the sensitivity study in Figure 2(c): DSI should help most when checkpoint errors for the target functional decorrelate across the trajectory, and should plateau when $K_{\mathrm{eff},\psi}$ saturates or the bias floor dominates.

5 Empirical Evaluation

We follow the data construction, market split, benchmark trading strategies, and evaluation protocol of Cont et al. (2025), assessing VaR and ES on the profit-and-loss distributions induced by $J = 65$

strategies. All generative methods use a fixed sampling budget of $N = 1000$ trajectories. DSI allocates this budget uniformly across $K = 20$ checkpoints selected with stride $M = 2$ from the corresponding training trajectory, without modifying the underlying architecture or objective. We compare against **WGAN** (Arjovsky et al., 2017) with gradient penalty (Gulrajani et al., 2017), **TailGAN** (Cont et al., 2025), and single-checkpoint **DDPM (EMA)** (Ho et al., 2020). Unless stated otherwise, each experiment is repeated 10 times, with an independent training run and random seed for each learned method. Full experimental details are provided in Appendix C (Sections C.1–C.6).

5.1 Evaluation Metrics and Tail-Risk Backtests

For confidence level α, the average relative error (RE) of VaR and ES across all J strategies is

$$\mathrm{RE}(\alpha) = \frac{1}{2J}\sum_{j=1}^{J}\left(\frac{\left|\widehat{\mathrm{VaR}}_\alpha(\tilde{\mathcal{X}}_j) - \mathrm{VaR}_\alpha(\mathcal{X}_j)\right|}{\left|\mathrm{VaR}_\alpha(\mathcal{X}_j)\right|} + \frac{\left|\widehat{\mathrm{ES}}_\alpha(\tilde{\mathcal{X}}_j) - \mathrm{ES}_\alpha(\mathcal{X}_j)\right|}{\left|\mathrm{ES}_\alpha(\mathcal{X}_j)\right|}\right), \tag{10}$$

where $\mathcal{X}_j$ and $\tilde{\mathcal{X}}_j$ denote the PnL distributions for strategy j computed from reference and generated data, respectively.

We evaluate distributional consistency via statistical backtests standard in risk management, conducted on synthetic data over 100 trials using $N = 1000$ generated samples. The Kupiec coverage test (Kupiec, 1995) evaluates whether the realised frequency of VaR violations matches the nominal tail level α. For example, a calibrated 5% VaR should be breached approximately 5% of the time. Since VaR alone does not measure the severity of losses beyond the threshold, we also use the Fissler-Ziegel score-based test (Fissler et al., 2016), which jointly evaluates VaR and ES calibration through a strictly consistent scoring function. We report rejection rates, where lower values indicate better agreement with the target tail-risk behaviour.

5.2 Empirical Results

Accuracy and Stability of Tail-Risk Estimation. Table 1 reports tail-estimation error under the fixed sampling budget.

Table 1: Tail-risk estimation and backtesting at $\alpha = 0.05$ under the fixed budget of $N = 1000$ trajectories. DSI uses $K = 20$ checkpoints and stride $M = 2$. Lower is better.

	Average RE (%)		Synthetic rejection rate (%)	
Method	Synthetic	Market	Coverage	Score-based
HSM	3.4 ± 2.6	11.4 ± 2.3	17.19	0.00
SE(1000) / Oracle	3.1 ± 2.3	2.6 ± 1.1	16.32	0.00
WGAN	17.3 ± 2.2	29.9 ± 1.9	70.35	14.92
TailGAN	5.1 ± 1.4	12.1 ± 1.3	33.55	0.29
DDPM (EMA)	7.7 ± 2.5	22.3 ± 4.3	53.62	0.06
WGAN (DSI)	16.6 ± 1.8	28.5 ± 1.4	68.12	12.34
TailGAN (DSI)	5.0 ± 1.3	11.9 ± 1.2	32.87	0.25
DDPM (DSI)	$\mathbf{3.7 \pm 0.2}$	$\mathbf{9.6 \pm 1.2}$	**20.26**	**0.00**

Note. Following Cont et al. (2025), RE is the average absolute relative error of VaR and ES across the $J = 65$ benchmark strategies. RE entries report mean $\pm$ standard deviation over 10 independent repetitions. For learned methods, each repetition uses an independently trained model. Historical Simulation Method (HSM) uses VaR and ES computed from historical data as predictions for held-out data. In the combined reference row, SE(1000) is the sampling error obtained from 1000 draws from the true synthetic distribution. Oracle denotes the test-set sampling-error reference for market RE and the benchmark computed from the known synthetic data-generating process for the backtests. The final two columns report rejection rates averaged across strategies over 100 synthetic-data trials for the Kupiec coverage test and the Fissler–Ziegel joint VaR–ES score-based test.

On synthetic data, DSI reduces single-checkpoint DDPM error from 7.7 ± 2.5 to 3.7 ± 0.2, approaching the sampling-error reference SE(1000). On market data, it reduces the error from 22.3 ± 4.3 to 9.6 ± 1.2 and outperforms the explicitly tail-optimized TailGAN without modifying the underlying

objective. Appendix D.1 provides rank-frequency PnL plots (Figures 4 and 5) and sensitivity results across $\alpha \in \{0.01, 0.05, 0.10\}$ (Table 4), which show the same qualitative improvement pattern under the fixed-budget protocol.

Checkpoint-error regimes. Table 2 links the heterogeneous gains in Table 1 to the regimes of Corollary 1. Its raw squared integrated-tail quantities retain the strategy-specific PnL scale and therefore diagnose within-model error regimes, whereas Table 1 reports equally weighted, strategy-normalised absolute VaR and ES errors for cross-model comparison.

Table 2: Integrated ES checkpoint-error diagnostics at $\alpha = 5\%$, $K = 20$, and $M = 2$, averaged over $J = 65$ strategies and five seeds. The columns report the bias floor, fluctuation variance, directional agreement, and effective checkpoint ratio. These raw unnormalised quantities diagnose each model's error regime; Table 1 reports cross-model RE.

Model	b^2_{ES}	σ^2_{ES}	Sign consensus	$K_{\text{eff},ES}/K$
DDPM	7.87×10^{-10}	1.79×10^{-8}	0.580	0.947
TailGAN	3.14×10^{-6}	1.47×10^{-5}	0.747	0.052
WGAN	1.32×10^{-7}	3.40×10^{-8}	0.871	0.650

DDPM combines a small bias floor with a high effective-checkpoint ratio and is therefore variance-reducible, whereas TailGAN is correlation-limited and WGAN exhibits strong sign consensus. The finite-budget decomposition in Appendix D.2, Table 5, further shows that mixture error dominates the sampling contribution for both adversarial trajectories. Thus DSI depends on averageable tail variation along the checkpoint trajectory rather than on the model family. Appendix D.2, including Figure 6, provides the construction and full diagnostic analysis.

Backtesting and Structural Fidelity. Table 1 reports formal risk backtests. Standard diffusion fails the coverage test in 53.62% of trials, reflecting lower-tail miscalibration. DSI reduces this rejection rate to 20.26% (approaching the Oracle's 16.32%) and achieves 0.00% rejection on the joint VaR/ES Fissler-Ziegel test, indicating improved probabilistic calibration under the fixed budget.

Multivariate and Temporal Dependency. DSI also preserves multivariate and temporal dependence metrics. Analyzing absolute differences between generated and reference correlation matrices and temporal autocorrelations (Figure 2(a–b)), DSI retains the structural fidelity of the underlying diffusion model and outperforms GAN-based methods on these diagnostics. This suggests that sample-space ensembling can improve marginal tail estimates without visibly degrading learned multivariate geometries. Full results for correlation and autocorrelation matrices are reported in Appendix D.3 (Figures 7–10).

Sensitivity to Mixture Size and Stride. Finally, we empirically evaluate DSI's sensitivity to the checkpoint selection parameters introduced in Section 3.4, the mixture size K and sampling stride M. As shown in Figure 2(c) and Table 6 in Appendix D.4, risk-estimation error improves rapidly as K increases from 5 to roughly 50, after which performance plateaus under the fixed sampling budget. This behavior is consistent with Theorem 1: increasing K helps while it increases the effective checkpoint count, but gains plateau once K_{eff} saturates and finite-budget remainders become non-negligible. Larger strides improve performance when K is small by reducing checkpoint-error autocorrelation, see Figure 6 in Appendix D.2. As K grows and the mixture spans a larger portion of the trajectory, the marginal effect of stride diminishes. Appendix D.5 reports the mechanism-matched aggregation controls under the same fixed budget.

Mechanism-matched aggregation controls. To test whether DSI's gain is explained by generic ensembling, we compare it with EMA, SWA, SWAG, and an independently trained Deep Ensemble under the same evaluation budget. DSI obtains the lowest relative error on both datasets under this protocol. The full comparison and implementation details are reported in Appendix D.5, Table 7.

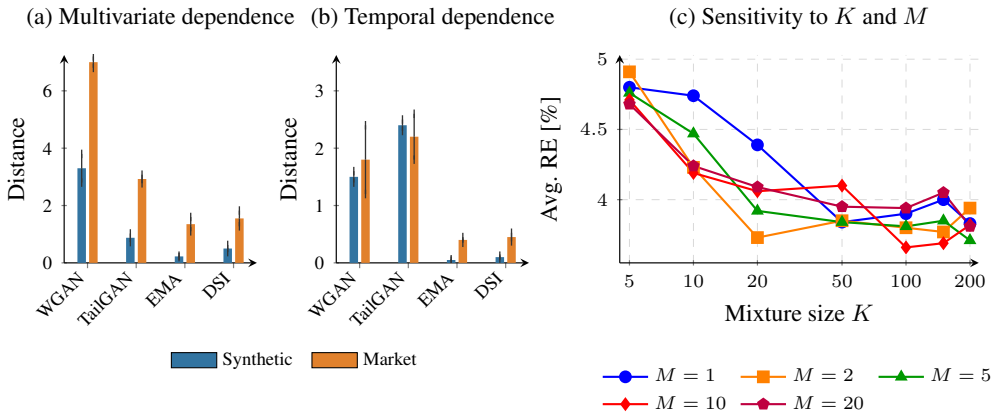


Figure 2: Structural fidelity and checkpoint-selection sensitivity under the fixed budget $N = 1000$. (a) Correlation-matrix distance and (b) autocorrelation distance between generated and reference data; bars show means and error bars show standard deviations (lower is better). (c) Average VaR/ES relative error at $\alpha = 5\%$ versus mixture size K for checkpoint strides M.

6 Discussion and Conclusion

We introduced Diachronic Sample Integration (DSI), a lightweight test-time sample-integration framework for finite-budget tail-risk estimation with generative simulators. By reusing stored checkpoints from a single stochastic training trajectory, DSI pools generated samples to target a trajectory-induced mixture distribution. Our theory and diagnostics show that DSI reduces checkpoint-fluctuation error when the trajectory contains averageable tail variation, while persistent bias or strong autocorrelation remains irreducible. Empirically, under fixed simulation budgets, DSI improves tail-risk estimation and calibration over TailGAN, a state-of-the-art tail-aware generative benchmark, without changing the generative objective.

Scope of DSI. DSI is not tied to a particular generative architecture, its effectiveness is governed by whether the post-burn-in checkpoint trajectory contains averageable tail variation. As formalized in Corollary 1 and Proposition 1, DSI helps in variance-reducible regimes but remains limited by persistent bias or strong checkpoint-error autocorrelation.

Broader impact. While our empirical evaluation rigorously focused on high-frequency financial time series, a canonical stress test due to their extreme signal-to-noise ratios, heavy tails, and complex multivariate dependencies, the mathematical formulation of DSI is entirely domain-agnostic. Theory derived in Section 4 apply to any setting where exact sampling is infeasible and downstream functionals depend on the extremes of a generative simulator. We anticipate DSI will be highly effective in other critical domains requiring robust extreme-value estimation under finite budgets, such as energy, climate modeling, structural engineering, and epidemiology.

Limitations and future work. Several limitations point to natural future work. First, DSI is primarily a variance-reduction mechanism over checkpoint fluctuations, it does not directly correct systematic tail bias. When all selected checkpoints share the same missing tail support or directional misspecification, the bias floor remains. Combining trajectory pooling with tail-guided or tail-constrained training methods is a promising direction for reducing this systematic component. Second, although our diagnostics characterize when DSI should work, they are currently evaluated after generating checkpoint samples. A useful next step is to develop lightweight compatibility diagnostics that can predict, early in training or with a small validation budget, whether a given model trajectory is variance-reducible, correlation-limited, or bias-limited. Third, our current estimator uses simple equal-weighted checkpoint mixtures, validation-weighted mixtures, adaptive stride selection, and bias-aware checkpoint selection may further improve finite-budget performance.

AI USE STATEMENT

In this work, we used generative AI tools for Generate synthetic data sets, assist with design or provide feedback on research methodology or experiments, assist in the writing of proofs, translation, clean and reformat dataset, support qualitative and thematic data analysis. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, implement methods, and interpret results. Additionally, we used generative AI tools for create or modify scientific figures or images, create or edit software code and edit a research paper to improve readability. We have reviewed all AI-assisted work. AI-generated figure layouts and code snippets were manually verified for correctness and consistency with the paper's methodology. All experimental code was tested to ensure reproducibility of reported results. AI-assisted text edits were reviewed for technical accuracy and alignment with our theoretical framework and empirical findings. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Source code for implementation and all experiments is available at `https://anonymous.4open.science/r/DSI-Submission/`. The synthetic data-generating process, statistical tests, and dependency analysis follow TailGAN Cont et al. (2025) and use their repository at `https://github.com/chaozhang-ox/Tail-GAN/` and the market portfolio experiments uses data from Nasdaq ITCH.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pp. 214–223. PMLR, 2017.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Basel Committee on Banking Supervision. International Convergence of Capital Measurement and Capital Standards: A Revised Framework, June 2006. URL `https://www.bis.org/publ/bcbs128.htm`.

Basel Committee on Banking Supervision. Minimum Capital Requirements for Market Risk, January 2019. URL `https://www.bis.org/bcbs/publ/d457.htm`.

Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. Supervisory guidance on model risk management. Federal Reserve SR Letter 11-7, Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, April 2011. URL `https://www.federalreserve.gov/boarddocs/srletters/2011/sr1107a1.pdf`. Issued jointly as OCC Bulletin 2011-12.

Rama Cont, Mihai Cucuringu, Renyuan Xu, and Chao Zhang. Tail-GAN: Learning to simulate tail risk scenarios. *Management Science*, 2025. doi: 10.1287/mnsc.2023.00936.

Pranoy Das, Kexin Fu, Abolfazl Hashemi, and Vijay Gupta. Principled long-tailed generative modeling via diffusion models. In *Advances in Neural Information Processing Systems*, 2025.

European Central Bank. ECB guide to internal models. Technical report, European Central Bank, June 2026. URL `https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.supervisory_guide_egim_202606.en.pdf`. Release 4.1.

Tobias Fissler, Johanna F. Ziegel, and Tilmann Gneiting. Expected shortfall is jointly elicitable with value at risk – implications for backtesting. *Risk*, pp. 58–61, January 2016.

Asadullah Hill Galib, Pang-Ning Tan, and Lifeng Luo. Fide: Frequency-inflated conditional diffusion model for extreme-aware time series generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 8803–8812, 2018.

Tal Gonen, Itai Pemper, Ilan Naiman, Nimrod Berman, and Omri Azencot. Time series generation under data scarcity: A unified generative modeling approach. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *International Conference on Learning Representations (ICLR)*, 2017.

Hongbin Huang, Minghua Chen, and Xiao Qiao. Fts-diffusion: Generative learning for financial time series with irregular and scale-invariant patterns. In *International Conference on Learning Representations (ICLR)*, 2024.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

Marcel Kollovieh, Abdul Fatir Ansari, Michael Bohlke-Schneider, Jasper Zschiegner, Hao Wang, and Yuyang Bernie Wang. Predict, refine, synthesize: Self-guiding diffusion models for probabilistic time series forecasting. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:28341–28364, 2023.

Paul H. Kupiec. Techniques for verifying the accuracy of risk measurement models. *The Journal of Derivatives*, 3(2):73–84, 1995. doi: 10.3905/jod.1995.407942. URL https://doi.org/10.3905/jod.1995.407942.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Junjie Li, Yang Liu, Weiqing Liu, Shikai Fang, Lewen Wang, Chang Xu, and Jiang Bian. Mars: a financial market simulation engine powered by generative foundation model. In *International Conference on Learning Representations (ICLR)*, 2025.

Jingwei Liu, Ling Yang, Hongyan Li, and Shenda Hong. Retrieval-augmented diffusion models for time series forecasting. *Advances in Neural Information Processing Systems (NeurIPS)*, 37: 2766–2786, 2024.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-tailed diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Tomonori Takahashi and Takayuki Mizuno. Generation of synthetic financial time series by diffusion models. *Quantitative Finance*, 2025.

Yuki Tanaka, Ryuji Hashimoto, Takehiro Takayanagi, Zhe Piao, Yuri Murayama, and Kiyoshi Izumi. Cofindiff: Controllable financial diffusion model for time series generation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2025.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*, pp. 23965–23998. PMLR, 2022.

Haochong Xia, Shuo Sun, Xinrun Wang, and Bo An. Market-gan: Adding control to financial market data generation with semantic context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38 of *14*, pp. 15996–16004, 2024.

Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yifeng Yu and Lu Yu. Diffusion models with heavy-tailed targets: Score estimation and sampling guarantees. *arXiv preprint arXiv:2601.06715*, 2026.

Xinyu Yuan and Yan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. In *International Conference on Learning Representations (ICLR)*, 2024.

CONTENTS

1 Introduction **1**

2 Related Work **2**

2.1 Generative Simulators and Tail-Sensitive Objectives 2

2.2 Model Ensembling and Averaging 3

3 Problem Formulation **3**

3.1 Probability Space and Tail-Dependent Functionals 3

3.2 Generative Approximation and the Single-Checkpoint Bottleneck 4

3.3 Checkpoint Ensembling on Generated Samples 4

3.4 Checkpoint Subsampling and Stride 5

4 Theoretical Analysis **5**

5 Empirical Evaluation **6**

5.1 Evaluation Metrics and Tail-Risk Backtests 7

5.2 Empirical Results 7

6 Discussion and Conclusion **9**

A DSI Procedure **14**

A.1 DSI Sampling Procedure 14

A.2 Checkpoint Selection and Burn-In 14

B Theoretical Details and Proofs **15**

B.1 Assumptions 15

B.2 Full Statements of the Main Results 16

B.3 Proof of Theorem 1 17

B.4 Proof of Corollary 1 21

B.5 Proof of Proposition 1 21

C Experimental Protocol and Implementation **22**

C.1 Evaluation Protocol and Fixed-Budget Design 22

C.2 Synthetic Data 22

C.3 Market Data 23

C.4 Baselines and Reference Estimators 23

C.5 Benchmark Trading Strategies 24

C.6 DDPM Architecture and Training 25

D Additional Empirical Analyses **25**

D.1 Tail-Accuracy Robustness 26
D.2 Checkpoint-Error Regime Diagnostics 29
D.3 Structural Fidelity 30
D.4 Mixture-Size and Checkpoint-Stride Sensitivity 32
D.5 Mechanism-Matched Aggregation Controls 32

A DSI PROCEDURE

This appendix formalizes the test-time sample aggregation procedure introduced in Sections 3.3 and 3.4, together with the checkpoint burn-in rule used in the experiments.

A.1 DSI SAMPLING PROCEDURE

Algorithm 1: Diachronic Sample Integration (DSI) Checkpoint-Mixture Sampling

Input: Set of model checkpoints along a training trajectory $\{\theta_t\}_{t=1}^{T}$, checkpoint count K, temporal stride M, total simulation budget N.
Output: Pooled synthetic dataset $\mathcal{X}$ representing samples from $\bar{P}_S$.
Identify start index t_0 after the burn-in phase;
$S \leftarrow \{t_0, t_0 + M, t_0 + 2M, \ldots, t_0 + (K-1)M\}$;
Discard any indices in S where $t > T$;
$K_S \leftarrow |S|$;
Set per-checkpoint sample size $m \leftarrow \lfloor N/K_S \rfloor$;
$\mathcal{X} \leftarrow \emptyset$;
foreach $t \in S$ **do**
  Draw m independent samples $\{X_{t,1}, \ldots, X_{t,m}\} \sim P_{\theta_t}$;
  $\mathcal{X} \leftarrow \mathcal{X} \cup \{X_{t,1}, \ldots, X_{t,m}\}$;
return $\mathcal{X}$.

A.2 CHECKPOINT SELECTION AND BURN-IN

The checkpoint set $S(M, K)$ defined in Algorithm 1 begins after a fixed burn-in period to avoid including under-trained models whose generative distributions deviate substantially from the target distribution. Specifically, models were trained for 3000 epochs, and checkpoints eligible for DSI were drawn uniformly from epoch 1000 onward. This choice coincides with the regime where the training loss and out-of-sample error have stabilized (Figure 3) and does not rely on any tail-risk metric, validation risk estimate, or oracle information. All reported DSI results use the same burn-in definition across models, datasets, and runs.

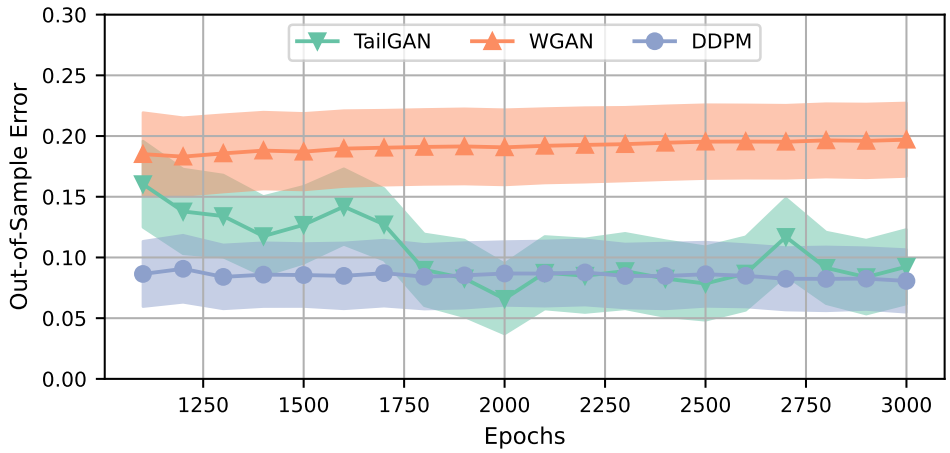


Figure 3: Post-burn-in out-of-sample error dynamics across training checkpoints. The solid line shows average relative error at $\alpha = 5\%$, and the shaded region denotes one standard deviation. DDPM maintains a low, stable error floor, TailGAN exhibits higher volatility, and WGAN shows an increasing error trend over later checkpoints.

B Theoretical Details and Proofs

This appendix provides the full assumptions, formal statements, and proofs for the finite-budget theory in Section 4. The proof uses a two-stage decomposition: first compare the checkpoint mixture $\bar{P}_S$ with the truth $P^\star$, and then compare the empirical DSI estimator with the fixed mixture $\bar{P}_S$. This avoids assuming a direct empirical plug-in linearization around $P^\star$.

B.1 Assumptions

Assumption 1 (Local tail regularity) *$F^\star$ has a unique α-quantile $q^\star$ and is continuously differentiable on $I_\delta = [q^\star - \delta, q^\star + \delta]$, with density $f^\star(q^\star) > 0$ and $\inf_{x \in I_\delta} f^\star(x) \geq c > 0$. When second-order remainders are invoked, $f^\star$ is locally Lipschitz on I_δ.*

Assumption 2 (Post-burn-in checkpoint-error process) *Let $X_t \sim P_t$ and $X^\star \sim P^\star$. Define $e_t(x) = F_t(x) - F^\star(x)$,*

$$U_t^q = e_t(q^\star), \tag{11}$$

$$U_t^{ES} = \int_{-\infty}^{q^\star} e_t(x)\, dx = \mathbb{E}[(q^\star - X_t)_+] - \mathbb{E}[(q^\star - X^\star)_+]. \tag{12}$$

After burn-in, assume $U_t^q = b_q + \xi_t^q$ and $U_t^{ES} = b_{ES} + \xi_t^{ES}$, where b_q, b_{ES} are persistent bias terms and ξ_t^q, ξ_t^{ES} are mean-zero weakly stationary fluctuations with variances $\sigma_q^2, \sigma_{ES}^2$ and autocorrelations $\rho_q(h), \rho_{ES}(h)$.

Assumption 3 (Sampling and lower-tail moments) *Assume $K_S \leq N$ and set $m = \lfloor N/K_S \rfloor$ and $N_S = K_S m$. Conditional on the selected checkpoint distributions $\{P_t : t \in S\}$, the generated samples $\{X_{t,i} : t \in S, 1 \leq i \leq m\}$ are independent with $X_{t,i} \sim P_t$. For all quantiles appearing below, the lower-tail variables $(q - X_t)_+$ have uniformly bounded second moments, and the analogous condition holds under $P^\star$.*

Assumption 4 (Local mixture regularity) *Let*

$$q_S := q_\alpha(\bar{P}_S), \tag{13}$$

$$\bar{e}_S(x) := \bar{F}_S(x) - F^\star(x), \tag{14}$$

$$\eta_S := \sup_{x \in I_\delta} |\bar{e}_S(x)|. \tag{15}$$

Define the two leading checkpoint-mixture errors

$$\bar{U}_{q,S} := \bar{F}_S(q^\star) - F^\star(q^\star), \tag{16}$$

$$\bar{U}_{ES,S} := \int_{-\infty}^{q^\star} (\bar{F}_S(x) - F^\star(x))\, dx, \tag{17}$$

and the local oscillation of the mixture error around $q^\star$,

$$\omega_S(r) := \sup_{\substack{x \in I_\delta \\ |x-q^\star| \le r}} |\bar{e}_S(x) - \bar{e}_S(q^\star)|. \tag{18}$$

Along the selected post-burn-in windows, the mixture remains in a local perturbation neighbourhood of $P^\star$: $q_S \in I_\delta$, $\bar{F}_S(q_S) = \alpha$, and $\bar{F}_S$ has a density $\bar{f}_S$ in a neighbourhood of q_S with $\bar{f}_S(q_S) \ge c_0 > 0$. The local oscillation and second-order terms are negligible relative to the leading checkpoint-mixture errors:

$$\mathbb{E}\big[(\omega_S(\eta_S/c) + \eta_S^2)^2\big] + \mathbb{E}\big[|\bar{U}_{q,S}|(\omega_S(\eta_S/c) + \eta_S^2)\big] = o\big(\mathbb{E}[\bar{U}_{q,S}^2]\big)\,, \tag{19}$$

and

$$\mathbb{E}[\eta_S^4] + \mathbb{E}[|\bar{U}_{ES,S}|\eta_S^2] = o\big(\mathbb{E}[\bar{U}_{ES,S}^2]\big)\,. \tag{20}$$

Assumption 4 is a local perturbation condition: it requires the selected post-burn-in mixture to remain in a neighbourhood where VaR and ES admit stable first-order expansions.

Assumption 5 (Standard stratified plug-in linearization) *Conditional on the selected checkpoint distributions, the empirical DSI estimators satisfy the usual Bahadur/plug-in linearizations around the fixed mixture $\bar{P}_S$:*

$$\widehat{q}^{DSI}_{\alpha,S,N} - q_S = -\frac{\widehat{F}_{S,N}(q_S) - \alpha}{\bar{f}_S(q_S)} + r^{sam}_{q,S,N}, \tag{21}$$

$$\widehat{\mathrm{ES}}^{DSI}_{\alpha,S,N} - \mathrm{ES}_\alpha(\bar{P}_S) = -\frac{1}{\alpha}\int_{-\infty}^{q_S} (\widehat{F}_{S,N}(x) - \bar{F}_S(x))\, dx + r^{sam}_{ES,S,N}, \tag{22}$$

with $\mathbb{E}[(r^{sam}_{q,S,N})^2 \mid \{P_t\}] = o(N_S^{-1})$ and $\mathbb{E}[(r^{sam}_{ES,S,N})^2 \mid \{P_t\}] = o(N_S^{-1})$ along the finite-budget sequences considered. These conditions hold under standard local density and moment assumptions for independent, stratified triangular arrays with positive density at the target mixture quantile. They are stated explicitly to separate sampling regularity from trajectory regularity in Assumption 4.

B.2 Full Statements of the Main Results

For completeness, we restate the three results from Section 4 in the order in which they appear there. The first result is the full version of Theorem 1, rather than a separate theorem.

For an autocorrelation function ρ, define

$$\kappa_\rho(K_S, M) = 1 + 2\sum_{\ell=1}^{K_S-1}\left(1 - \frac{\ell}{K_S}\right)\rho(\ell M), \tag{23}$$

$$K_{\mathrm{eff},\rho}(K_S, M) = \frac{K_S}{\kappa_\rho(K_S, M)}, \tag{24}$$

assuming $\kappa_\rho(K_S, M) > 0$. Define the conditional sampling variances at the mixture quantile

$$A^{\mathrm{mix}}_{q,N}(S) = \frac{1}{K_S^2 m}\sum_{t\in S} F_t(q_S)(1 - F_t(q_S)), \tag{25}$$

$$A^{\mathrm{mix}}_{ES,N}(S) = \frac{1}{K_S^2 m}\sum_{t\in S} \mathrm{Var}_{P_t}[(q_S - X_t)_+]. \tag{26}$$

For comparison, write $A_{q,N}(S)$ and $A_{ES,N}(S)$ for the same two quantities with q_S replaced by $q^\star$. Unless explicitly conditioned, expectations below are taken over both the post-burn-in checkpoint-error process and the synthetic samples drawn conditionally from the selected checkpoints.

Theorem 1 (Two-stage finite-budget DSI MSE expansion, full statement). *Suppose Assumptions 1-5 hold. Define*

$$L_q^{traj} = b_q^2 + \frac{\sigma_q^2}{K_{\text{eff},q}(K_S, M)}, \qquad L_{q,N}^{sam} = \mathbb{E}\left[\frac{A_{q,N}^{\text{mix}}(S)}{\bar{f}_S(q_S)^2}\right], \tag{27}$$

$$L_{ES}^{traj} = b_{ES}^2 + \frac{\sigma_{ES}^2}{K_{\text{eff},ES}(K_S, M)}, \qquad L_{ES,N}^{sam} = \mathbb{E}A_{ES,N}^{\text{mix}}(S). \tag{28}$$

Then

$$\mathbb{E}\left[\left(\widehat{q}_{\alpha,S,N}^{DSI} - q^\star\right)^2\right] = \frac{L_q^{traj}}{f^\star(q^\star)^2} + L_{q,N}^{sam} + \mathcal{R}_{q,N}, \tag{29}$$

$$\mathbb{E}\left[\left(\widehat{\text{ES}}_{\alpha,S,N}^{DSI} - \text{ES}_\alpha(P^\star)\right)^2\right] = \frac{(L_{ES}^{traj} + L_{ES,N}^{sam})}{\alpha^2} + \mathcal{R}_{ES,N}, \tag{30}$$

where

$$\mathcal{R}_{q,N} = o\left(\frac{L_q^{traj}}{f^\star(q^\star)^2}\right) + o(N_S^{-1}), \tag{31}$$

$$\mathcal{R}_{ES,N} = o\left(\frac{L_{ES}^{traj}}{\alpha^2}\right) + o(N_S^{-1}). \tag{32}$$

If, additionally, $\bar{f}_S(q_S) = f^\star(q^\star) + o_{\mathbb{P}}(1)$ and the stratified variances at q_S are first-order equivalent to those at $q^\star$, this reduces to the simpler common-denominator form with $A_{q,N}(S)$ and $A_{ES,N}(S)$ evaluated at $q^\star$. In the notation of Theorem 1, $L_{q,N}^{sam} = L_{q,N}^{sam}$ and $L_{ES,N}^{sam} = L_{ES,N}^{sam}/\alpha^2$.

Corollary 1 (Mechanistic regimes) *Theorem 1 identifies three regimes: (i) a* variance-reducible *regime where the corresponding bias term is small and K_{eff} grows, (ii) a* correlation-limited *regime where checkpoint fluctuations are present but highly autocorrelated, so $K_{\text{eff}} \ll K_S$, and (iii) a* bias-limited *regime where b^2 dominates and DSI cannot remove the leading error term.*

Proposition 1 (Failure modes) *DSI cannot correct perfectly correlated checkpoint errors, shared directional lower-tail misspecification across all selected checkpoints, or persistent bias floors. In particular, if $\rho(\ell M) = 1$ for all $\ell = 1, \ldots, K_S - 1$, then $K_{\text{eff}} = 1$. If, for some sign $s \in \{-1, 1\}$ and nonnegative function a, $s\{F_t(x) - F^\star(x)\} \geq a(x)$ for all $t \in S$ on a tail interval, then $s\{\bar{F}_S(x) - F^\star(x)\} \geq a(x)$ on that interval. Even when $K_{\text{eff}} \to \infty$ and $N \to \infty$, the first-order VaR and ES MSE floors remain $b_q^2/f^\star(q^\star)^2$ and b_{ES}^2/α^2, respectively.*

B.3 Proof of Theorem 1

The proof uses three preparatory results, which we state and prove in the order in which they are used. Lemma 1 gives the local first-order expansions of VaR and ES, Proposition 2 characterises the reduction in checkpoint error under trajectory averaging, and Proposition 3 gives the conditional sampling identities for the fixed checkpoint mixture. We then combine these ingredients to complete the two-stage MSE expansion.

B.3.1 Tail-Functional Expansion

Lemma 1 (Smooth first-order tail-functional expansions) *Let G be a continuous CDF with finite lower-tail first moment, $q_G = q_\alpha(G) \in I_\delta$, and $G(q_G) = \alpha$. Write $g = G - F^\star$, $\Delta = q_G - q^\star$, $\eta = \sup_{x \in I_\delta} |g(x)|$, and*

$$\omega_G(r) := \sup_{\substack{x \in I_\delta \\ |x - q^\star| \leq r}} |g(x) - g(q^\star)|. \tag{33}$$

Under Assumption 1, $|\Delta| \leq \eta/c$ and

$$q_\alpha(G) - q^\star = -\frac{g(q^\star)}{f^\star(q^\star)} + O\big(\omega_G(\eta/c) + \eta^2\big), \tag{34}$$

while

$$\mathrm{ES}_\alpha(G) - \mathrm{ES}_\alpha(P^\star) = -\frac{1}{\alpha}\int_{-\infty}^{q^\star} g(x)\,dx + O(\eta^2). \tag{35}$$

Proof. Because $G(q_G) = \alpha$ and $F^\star(q^\star) = \alpha$,

$$0 = G(q_G) - F^\star(q^\star) = F^\star(q^\star + \Delta) - F^\star(q^\star) + g(q_G). \tag{36}$$

The lower density bound in Assumption 1 implies

$$c|\Delta| \le |F^\star(q^\star + \Delta) - F^\star(q^\star)| = |g(q_G)| \le \eta, \tag{37}$$

so $|\Delta| \le \eta/c$. A Taylor expansion of $F^\star$ around $q^\star$ gives

$$0 = f^\star(q^\star)\Delta + g(q^\star) + \{g(q_G) - g(q^\star)\} + O(\Delta^2). \tag{38}$$

Since $|\Delta| \le \eta/c$, the oscillation term is bounded by $\omega_G(\eta/c)$ and $\Delta^2 = O(\eta^2)$. Hence

$$\Delta = -\frac{g(q^\star)}{f^\star(q^\star)} + O\big(\omega_G(\eta/c) + \eta^2\big), \tag{39}$$

which proves the VaR expansion.

For ES, use the quantile-integral/lower-tail CDF identity, valid for continuous distributions with finite lower-tail first moment,

$$\mathrm{ES}_\alpha(G) = q_G - \frac{1}{\alpha}\int_{-\infty}^{q_G} G(x)\,dx, \tag{40}$$

$$\mathrm{ES}_\alpha(P^\star) = q^\star - \frac{1}{\alpha}\int_{-\infty}^{q^\star} F^\star(x)\,dx. \tag{41}$$

Subtracting gives

$$\mathrm{ES}_\alpha(G) - \mathrm{ES}_\alpha(P^\star) = \Delta - \frac{1}{\alpha}\left[\int_{-\infty}^{q^\star} g(x)\,dx + \int_{q^\star}^{q^\star+\Delta} G(x)\,dx\right]. \tag{42}$$

The second integral satisfies

$$\int_{q^\star}^{q^\star+\Delta} G(x)\,dx = \alpha\Delta + O(\Delta^2) + O(\eta|\Delta|), \tag{43}$$

where the last term bounds the perturbation contribution on I_δ. Since $|\Delta| \le \eta/c$, the Δ terms cancel and the remainder is $O(\eta^2)$:

$$\mathrm{ES}_\alpha(G) - \mathrm{ES}_\alpha(P^\star) = -\frac{1}{\alpha}\int_{-\infty}^{q^\star} g(x)\,dx + O(\eta^2). \tag{44}$$

■

B.3.2 Checkpoint-Error Averaging

Proposition 2 (Checkpoint-error averaging) *Under Assumption 2,*

$$\mathbb{E}\left[\left(\frac{1}{K_S}\sum_{t\in S} U_t^q\right)^2\right] = b_q^2 + \frac{\sigma_q^2}{K_{\mathrm{eff},q}(K_S, M)}, \tag{45}$$

and

$$\mathbb{E}\left[\left(\frac{1}{K_S}\sum_{t\in S} U_t^{ES}\right)^2\right] = b_{ES}^2 + \frac{\sigma_{ES}^2}{K_{\mathrm{eff},ES}(K_S, M)}. \tag{46}$$

Proof. We prove the VaR statement and the ES statement is identical. By Assumption 2,

$$U_t^q = b_q + \xi_t^q, \qquad \mathbb{E}[\xi_t^q] = 0. \tag{47}$$

Thus

$$\frac{1}{K_S}\sum_{t\in S} U_t^q = b_q + \frac{1}{K_S}\sum_{t\in S}\xi_t^q. \tag{48}$$

Squaring and taking expectations gives

$$\mathbb{E}\left[\left(\frac{1}{K_S}\sum_{t\in S} U_t^q\right)^2\right] = b_q^2 + \mathbb{E}\left[\left(\frac{1}{K_S}\sum_{t\in S}\xi_t^q\right)^2\right], \tag{49}$$

because the cross term vanishes. Writing $S = \{t_0 + rM : r = 0, \dots, K_S - 1\}$, weak stationarity gives

$$\mathrm{Cov}(\xi_{t_0+rM}^q, \xi_{t_0+sM}^q) = \sigma_q^2 \rho_q(|r-s|M). \tag{50}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{1}{K_S}\sum_{r=0}^{K_S-1}\xi_{t_0+rM}^q\right)^2\right] &= \frac{\sigma_q^2}{K_S^2}\sum_{r=0}^{K_S-1}\sum_{s=0}^{K_S-1}\rho_q(|r-s|M) \\
&= \frac{\sigma_q^2}{K_S}\left[1 + 2\sum_{\ell=1}^{K_S-1}\left(1-\frac{\ell}{K_S}\right)\rho_q(\ell M)\right] \\
&= \frac{\sigma_q^2}{K_{\mathrm{eff},q}(K_S, M)}.
\end{aligned} \tag{51}$$

This proves the VaR identity. Replacing U_t^q by U_t^{ES} proves the ES identity. ∎

B.3.3 Stratified DSI Sampling Identities

Proposition 3 (Stratified DSI sampling identities) *Under Assumption 3, condition on the selected checkpoint distributions. For any fixed $x \in \mathbb{R}$,*

$$\mathbb{E}[\widehat{F}_{S,N}(x) \mid \{P_t : t \in S\}] = \bar{F}_S(x), \tag{52}$$

and

$$\mathrm{Var}[\widehat{F}_{S,N}(x) \mid \{P_t : t \in S\}] = \frac{1}{K_S^2 m}\sum_{t\in S} F_t(x)(1-F_t(x)). \tag{53}$$

If $m = N/K_S$ exactly, then this variance is at most $\bar{F}_S(x)(1-\bar{F}_S(x))/N$. Moreover, for any fixed threshold q with finite lower-tail second moments,

$$Z_{ES,N}(q) := \int_{-\infty}^{q} (\widehat{F}_{S,N}(x) - \bar{F}_S(x))\, dx \tag{54}$$

satisfies

$$\mathbb{E}[Z_{ES,N}(q) \mid \{P_t : t \in S\}] = 0, \tag{55}$$

and

$$\mathrm{Var}[Z_{ES,N}(q) \mid \{P_t : t \in S\}] = \frac{1}{K_S^2 m}\sum_{t\in S}\mathrm{Var}_{P_t}[(q - X_t)_+]. \tag{56}$$

Proof. Condition on the selected checkpoint distributions. For any fixed x,

$$\widehat{F}_{S,N}(x) = \frac{1}{K_S}\sum_{t\in S}\widehat{F}_{t,m}(x), \tag{57}$$

$$\widehat{F}_{t,m}(x) = \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}\{X_{t,i} \le x\}. \tag{58}$$

Because $\mathbb{E}[\widehat{F}_{t,m}(x) \mid P_t] = F_t(x)$,

$$\mathbb{E}[\widehat{F}_{S,N}(x) \mid \{P_t : t \in S\}] = \frac{1}{K_S} \sum_{t \in S} F_t(x) = \bar{F}_S(x). \tag{59}$$

Conditional independence across samples and checkpoints gives

$$\mathrm{Var}[\widehat{F}_{S,N}(x) \mid \{P_t : t \in S\}] = \frac{1}{K_S^2} \sum_{t \in S} \mathrm{Var}[\widehat{F}_{t,m}(x) \mid P_t] = \frac{1}{K_S^2 m} \sum_{t \in S} F_t(x)(1 - F_t(x)). \tag{60}$$

If $m = N/K_S$, this becomes

$$\frac{1}{K_S N} \sum_{t \in S} F_t(x)(1 - F_t(x)). \tag{61}$$

Since $p \mapsto p(1-p)$ is concave on $[0, 1]$, Jensen's inequality yields

$$\frac{1}{K_S} \sum_{t \in S} F_t(x)(1 - F_t(x)) \le \bar{F}_S(x)(1 - \bar{F}_S(x)). \tag{62}$$

For the integrated identity, fix q. For any sample value X,

$$\int_{-\infty}^{q} \mathbf{1}\{X \le x\}\, dx = (q - X)_+. \tag{63}$$

Therefore

$$Z_{ES,N}(q) = \frac{1}{K_S} \sum_{t \in S} \left[\frac{1}{m} \sum_{i=1}^{m} (q - X_{t,i})_+ - \mathbb{E}_{P_t}(q - X_t)_+ \right]. \tag{64}$$

The conditional mean is zero, and conditional independence gives

$$\mathrm{Var}[Z_{ES,N}(q) \mid \{P_t : t \in S\}] = \frac{1}{K_S^2 m} \sum_{t \in S} \mathrm{Var}_{P_t}[(q - X_t)_+]. \tag{65}$$

■

B.3.4 Completion of the Proof

Proof. We prove VaR first. Decompose

$$\widehat{q}^{DSI}_{\alpha,S,N} - q^\star = \underbrace{(q_S - q^\star)}_{D_{2,q}\text{: trajectory}} + \underbrace{(\widehat{q}^{DSI}_{\alpha,S,N} - q_S)}_{D_{1,q}\text{: sampling}}. \tag{66}$$

For the trajectory term, Lemma 1 applied to $G = \bar{F}_S$ gives

$$D_{2,q} = -\frac{\bar{F}_S(q^\star) - F^\star(q^\star)}{f^\star(q^\star)} + r^{traj}_{q,S} = -\frac{1}{f^\star(q^\star)} \frac{1}{K_S} \sum_{t \in S} U^q_t + r^{traj}_{q,S}, \tag{67}$$

where the deterministic remainder is negligible in the sense of Assumption 4. Proposition 2 then yields

$$\mathbb{E}[D^2_{2,q}] = \frac{1}{f^\star(q^\star)^2} \left(b_q^2 + \frac{\sigma_q^2}{K_{\mathrm{eff},q}(K_S, M)} \right) + o(L^{traj}_q). \tag{68}$$

For the sampling term, Assumption 5 gives

$$D_{1,q} = -\frac{Z_{q,N}}{\bar{f}_S(q_S)} + r^{sam}_{q,S,N}, \tag{69}$$

$$Z_{q,N} := \widehat{F}_{S,N}(q_S) - \alpha. \tag{70}$$

Conditional on the trajectory, $\mathbb{E}[Z_{q,N} \mid \{P_t\}] = 0$ because $\bar{F}_S(q_S) = \alpha$, and Proposition 3 gives

$$\mathrm{Var}[Z_{q,N} \mid \{P_t\}] = A^{\mathrm{mix}}_{q,N}(S). \tag{71}$$

Therefore

$$\mathbb{E}[D_{1,q}^2] = \mathbb{E}\left[\frac{A_{q,N}^{\mathrm{mix}}(S)}{\bar{f}_S(q_S)^2}\right] + o(N_S^{-1}) = L_{q,N}^{sam} + o(N_S^{-1}). \tag{72}$$

The leading sampling term has conditional mean zero, so

$$\mathbb{E}[D_{1,q}D_{2,q}] = \mathbb{E}\big[D_{2,q}\,\mathbb{E}(r_{q,S,N}^{sam} \mid \{P_t\})\big]. \tag{73}$$

By Cauchy-Schwarz, Assumption 4, and Appendix Assumption 5, this cross term is negligible relative to $L_q^{traj} + L_{q,N}^{sam}$. Combining the three pieces proves the VaR expansion.

For ES, use the analogous decomposition

$$\widehat{\mathrm{ES}}_{\alpha,S,N}^{DSI} - \mathrm{ES}_\alpha(P^\star) = \underbrace{(\mathrm{ES}_\alpha(\bar{P}_S) - \mathrm{ES}_\alpha(P^\star))}_{D_{2,ES}} + \underbrace{(\widehat{\mathrm{ES}}_{\alpha,S,N}^{DSI} - \mathrm{ES}_\alpha(\bar{P}_S))}_{D_{1,ES}}. \tag{74}$$

Lemma 1 gives

$$D_{2,ES} = -\frac{1}{\alpha}\int_{-\infty}^{q^\star} (\bar{F}_S(x) - F^\star(x))\,dx + r_{ES,S}^{traj} = -\frac{1}{\alpha}\frac{1}{K_S}\sum_{t\in S} U_t^{ES} + r_{ES,S}^{traj}, \tag{75}$$

where the deterministic remainder is negligible in the sense of Assumption 4. Proposition 2 yields

$$\mathbb{E}[D_{2,ES}^2] = \frac{1}{\alpha^2}\left(b_{ES}^2 + \frac{\sigma_{ES}^2}{K_{\mathrm{eff},ES}(K_S, M)}\right) + o(L_{ES}^{traj}). \tag{76}$$

For the sampling term, Assumption 5 gives

$$D_{1,ES} = -\frac{1}{\alpha}Z_{ES,N}(q_S) + r_{ES,S,N}^{sam}, \tag{77}$$

$$Z_{ES,N}(q_S) := \int_{-\infty}^{q_S} (\widehat{F}_{S,N}(x) - \bar{F}_S(x))\,dx. \tag{78}$$

By Proposition 3, conditional on the trajectory,

$$\mathbb{E}[Z_{ES,N}(q_S) \mid \{P_t\}] = 0, \tag{79}$$

$$\mathrm{Var}[Z_{ES,N}(q_S) \mid \{P_t\}] = A_{ES,N}^{\mathrm{mix}}(S). \tag{80}$$

Thus

$$\mathbb{E}[D_{1,ES}^2] = \frac{1}{\alpha^2}\mathbb{E}A_{ES,N}^{\mathrm{mix}}(S) + o(N_S^{-1}). \tag{81}$$

The cross term is again negligible because the leading sampling term has conditional mean zero, the population remainder is controlled by Assumption 4, and the sampling remainder is mean-square negligible by Appendix Assumption 5. Combining trajectory, sampling, and cross terms proves the ES expansion. ∎

B.4 Proof of Corollary 1

Proof. The regimes follow directly from Theorem 1. If b^2 is small and K_{eff} grows, the checkpoint-fluctuation term $\sigma^2/K_{\mathrm{eff}}$ decreases, producing a variance-reducible regime. If $K_{\mathrm{eff}} \ll K_S$ because autocorrelation is persistent, the variance term does not decrease much, producing a correlation-limited regime. If b^2 dominates the other terms, increasing K_{eff} or N cannot remove the dominant error, producing a bias-limited regime. ∎

B.5 Proof of Proposition 1

Proof. If $\rho(\ell M) = 1$ for every $\ell = 1, \ldots, K_S - 1$, then

$$\begin{aligned}\kappa_\rho(K_S, M) &= 1 + 2\sum_{\ell=1}^{K_S-1}\left(1 - \frac{\ell}{K_S}\right) \\ &= 1 + 2\left[(K_S - 1) - \frac{1}{K_S}\frac{(K_S-1)K_S}{2}\right] = K_S.\end{aligned} \tag{82}$$

Thus $K_{\text{eff}} = K_S/K_S = 1$.

For shared directional tail misspecification, suppose that for some $s \in \{-1, 1\}$,

$$s\{F_t(x) - F^\star(x)\} \geq a(x) \qquad \text{for every } t \in S. \tag{83}$$

Then by linearity of the mixture CDF,

$$\begin{aligned} s\{\bar{F}_S(x) - F^\star(x)\} &= \frac{1}{K_S}\sum_{t\in S} s\{F_t(x) - F^\star(x)\} \\ &\geq \frac{1}{K_S}\sum_{t\in S} a(x) = a(x). \end{aligned} \tag{84}$$

Thus the mixture inherits any common directional lower-tail error, including both undercoverage and overcoverage.

Finally, the bias floors follow by taking $K_{\text{eff}} \to \infty$ and $N \to \infty$ in Theorem 1. The checkpoint-fluctuation and finite-sampling terms vanish, leaving $b_q^2/f^\star(q^\star)^2$ for VaR and b_{ES}^2/α^2 for ES, up to negligible first-order remainders. ∎

C Experimental Protocol and Implementation

Our theoretical analysis decomposes finite-budget tail-risk error under the trajectory-induced mixture $\bar{P}_S$ into trajectory bias, checkpoint-fluctuation variance, and synthetic-sampling variance. We evaluate DSI in this regime to assess (i) stabilization of tail-functional estimation relative to single-checkpoint inference, (ii) preservation of structural properties (cross-asset and temporal dependence), and (iii) robustness to the checkpoint-selection parameters, the mixture size K and stride M.

Following the experimental protocol of Cont et al. (2025), tail-dependent functionals are evaluated on profit-and-loss (PnL) distributions induced by J benchmark trading strategies applied to multivariate return processes. The univariate functionals q_α and ES_α analysed in Section 4 are thus computed on strategy-level empirical distributions.

C.1 Evaluation Protocol and Fixed-Budget Design

To isolate estimator variance and bias, all generative methods are evaluated under a fixed sampling budget of $N = 1000$ generated trajectories, reflecting practical risk-management constraints. For DSI, drawing from the selected checkpoint set S implies $m = \lfloor N/K_S \rfloor$ samples per checkpoint, directly matching the stratified finite-budget analysis in Theorem 1. Unless stated otherwise, each experiment is repeated 10 times, with an independent training run and random seed for each learned method, and DSI uses mixture size $K = 20$ and stride $M = 2$. Appendix A gives the full sampling procedure and burn-in rule.

C.2 Synthetic Data

Following Cont et al. (2025), we construct a five-dimensional return process $\mathbf{X}(t)$ combining heterogeneous components (Gaussian, AR(1), and heavy-tailed Student-t GARCH(1,1)) with fixed multivariate correlation. Because the true data-generating distribution $P^\star$ is analytically known, exact oracle values of $q_\alpha(P^\star)$ and $\mathrm{ES}_\alpha(P^\star)$ are available for direct error evaluation.

We follow the synthetic data-generation procedure of Cont et al. (2025) exactly to ensure a fair and controlled comparison across methods. For each time step $t \in \{1, \ldots, T\}$, we first sample a latent innovation vector

$$\mathbf{u}_t = (u_{1,t}, \ldots, u_{5,t})^\top \sim \mathcal{N}(\mathbf{0}, \Sigma), \tag{85}$$

where $\Sigma \in \mathbb{R}^{5\times 5}$ is a covariance matrix. In addition, we sample independent chi-square random variables $v_{1,t} \sim \chi^2(\nu_1)$ and $v_{2,t} \sim \chi^2(\nu_2)$, which are independent of $\mathbf{u}_t$.

The five asset price increments are then generated according to

$$\begin{aligned}
\Delta p_{1,t} &= u_{1,t}, \\
\Delta p_{2,t} &= \phi_1 \Delta p_{2,t-1} + u_{2,t}, \\
\Delta p_{3,t} &= \phi_2 \Delta p_{3,t-1} + u_{3,t}, \\
\Delta p_{4,t} &= \varepsilon_{4,t} = \sigma_{4,t}\eta_{1,t}, \\
\Delta p_{5,t} &= \varepsilon_{5,t} = \sigma_{5,t}\eta_{2,t},
\end{aligned}$$

where the conditional variances evolve according to GARCH(1,1) dynamics:

$$\begin{aligned}
\sigma^2_{4,t} &= \gamma_4 + \kappa_4 \varepsilon^2_{4,t-1} + \beta_4 \sigma^2_{4,t-1}, \\
\sigma^2_{5,t} &= \gamma_5 + \kappa_5 \varepsilon^2_{5,t-1} + \beta_5 \sigma^2_{5,t-1}.
\end{aligned}$$

The standardized innovations are defined as

$$\eta_{1,t} = \frac{u_{4,t}}{\sqrt{v_{1,t}/\nu_1}}, \qquad \eta_{2,t} = \frac{u_{5,t}}{\sqrt{v_{2,t}/\nu_2}}, \tag{86}$$

which induce heavy-tailed marginal behaviour.

We set $T = 100$, corresponding to the number of intraday observations within one trading day. A correlation matrix R is first generated by sampling each element independently from $\mathrm{Unif}(0, 1)$. Annualized standard deviations s_i are sampled uniformly from $[0.3, 0.5]$, and the covariance matrix Σ is constructed as

$$\Sigma_{ij} = \frac{s_i}{\sqrt{255T}} \frac{s_j}{\sqrt{255T}} R_{ij}, \qquad i, j = 1, \ldots, 5. \tag{87}$$

The remaining parameters are fixed as follows: $\phi_1 = 0.5$, $\phi_2 = -0.12$, $\nu_1 = 5$, $\nu_2 = 10$, $\kappa_4, \kappa_5 \sim \mathrm{Unif}(0.08, 0.12)$, $\beta_4, \beta_5 \sim \mathrm{Unif}(0.825, 0.875)$, and $\gamma_4, \gamma_5 \sim \mathrm{Unif}(0.03, 0.07)$. Main results are reported for the risk level $\alpha = 0.05$, while sensitivity results for $\alpha \in \{0.01, 0.05, 0.10\}$ are provided in Table 4.

C.3 Market Data

Following the experimental setup and asset-selection criteria established in Cont et al. (2025), we use the Nasdaq ITCH limit order book dataset for five liquid assets (AAPL, AMZN, GOOG, JPM, QQQ), sampled every $\Delta = 9$ seconds. Each trajectory spans 100 time steps (15 minutes). Training uses November 2019 data (6,300 samples), and testing uses the first week of December 2019 for out-of-sample evaluation. Table 3 reports the empirical VaR and ES estimates across the benchmark strategies.

	Static Buy-and-Hold		Mean-Reversion		Trend-Following	
Asset	VaR	ES	VaR	ES	VaR	ES
AAPL	-0.366	-0.591	-0.307	-0.491	-0.382	-0.527
AMZN	-0.420	-0.606	-0.347	-0.508	-0.330	-0.453
GOOG	-0.377	-0.572	-0.340	-0.510	-0.347	-0.462
JPM	-0.360	-0.519	-0.287	-0.405	-0.350	-0.473
QQQ	-0.223	-0.447	-0.192	-0.349	-0.191	-0.289

Table 3: Empirical Value-at-Risk (VaR) and Expected Shortfall (ES) on market training data for $\alpha = 5\%$.

C.4 Baselines and Reference Estimators

We compare DSI against general adversarial generation (**WGAN** (Arjovsky et al., 2017) with gradient penalty (Gulrajani et al., 2017)), explicitly tail-optimized generation (**TailGAN** (Cont et al., 2025)), and standard single-checkpoint diffusion inference (**DDPM (EMA)** (Ho et al., 2020)).

The DDPM (EMA) baseline generates samples using an exponential moving average of the network weights, a common practice in diffusion-model training that smooths high-frequency optimization

noise. We use the best EMA checkpoint as the single-checkpoint diffusion baseline and apply DSI to the same training trajectory without changing the architecture or objective. Section C.6 gives the diffusion architecture and training configuration.

Historical Simulation (**HSM**) is a nonparametric risk-estimation reference that uses VaR and ES computed from historical PnL values as predictions for held-out data, without fitting a generative model. On synthetic data, **SE(1000)** denotes the finite-sample error obtained by estimating VaR and ES from exactly 1000 draws from the true synthetic process. On market data, **Oracle** denotes the corresponding sampling error obtained from the empirical test distribution, with the full test set providing the reference risk measures. In the synthetic backtests, Oracle instead denotes the known-data-generating-process benchmark against which the estimated VaR and ES are tested.

C.5 Benchmark Trading Strategies

Following Cont et al. (2025), we evaluate downstream tail risk on PnL distributions induced by $J = 65$ benchmark trading strategies. These strategies form three functional groups: (i) **Buy-and-Hold**, which tests heavy-tailed multivariate dependence, (ii) **Mean Reversion**, which tests short-term negative autocorrelation, and (iii) **Trend Following**, which tests long-term momentum. Each strategy is a deterministic functional mapping a multivariate asset-price trajectory $\mathbf{S}_{1:T} = \{\mathbf{S}_t\}_{t=1}^{T} \in \mathbb{R}^{d\times T}$, derived from the generated return sequence, to a univariate portfolio PnL sequence.

The terminal PnL over the investment horizon instantiates the random variable X introduced in Section 3, upon which the statistical risk functionals $\mathrm{VaR}_\alpha(X)$ and $\mathrm{ES}_\alpha(X)$ are evaluated. Strategies are applied identically to empirical reference data and synthetic trajectories, using only information available up to time t.

C.5.1 Buy-and-Hold (Multivariate Dependence)

The Buy-and-Hold strategy represents a static allocation baseline. Because it aggregates simultaneous price movements without dynamic trading, it tests the generative model's ability to preserve heavy-tailed multivariate dependence. Given a fixed portfolio weight vector $\mathbf{w} \in \mathbb{R}^d$, derived from a predefined allocation matrix, the portfolio value V_t at time t is computed as

$$V_t = \mathbf{w}^\top \mathbf{S}_t = \sum_{i=1}^{d} w_i S_{i,t}. \tag{88}$$

The strategy takes an initial position and holds it over the entire investment horizon. The terminal PnL is $X = V_T - V_1$.

C.5.2 Mean Reversion (Short-Term Temporal Dependence)

The Mean Reversion (MR) strategy tests preservation of short-term negative autocorrelation, operating on the assumption that asset prices eventually revert to a historical anchor. For a given asset, we define the anchoring mean μ_{init} as the average price over an initial lookback window of length W:

$$\mu_{\text{init}} = \frac{1}{W}\sum_{k=1}^{W} S_k. \tag{89}$$

At each subsequent active time step $t > W$, a normalized deviation score z_t measures the divergence of the current price from this initial mean:

$$z_t = \frac{S_t - \mu_{\text{init}}}{c}, \tag{90}$$

where $c > 0$ is a fixed empirical scaling factor (e.g., $c = 0.01$). Trading signals are generated dynamically based on empirical quantile thresholds. The strategy initiates a **Long** position when $z_t < \tau_{\text{lower}}$, under the assumption that the asset is undervalued and will revert upward, and a **Short** position when $z_t > \tau_{\text{upper}}$, identifying overvaluation.

C.5.3 TREND FOLLOWING (LONG-TERM TEMPORAL DEPENDENCE)

The Trend Following (TF) strategy implements a Moving Average Crossover system to capture market momentum, thereby probing long-term positive autocorrelation in the generated paths. For a given asset, we compute two simple moving averages (SMA), a short-term average over a window of length W and a long-term average over a window of length $2W$:

$$\text{SMA}_{\text{short},t} = \frac{1}{W}\sum_{k=0}^{W-1} S_{t-k}, \quad \text{SMA}_{\text{long},t} = \frac{1}{2W}\sum_{k=0}^{2W-1} S_{t-k}. \tag{91}$$

The trading signal is derived from the normalized divergence between these two temporal smoothing filters:

$$z_t = \frac{\text{SMA}_{\text{short},t} - \text{SMA}_{\text{long},t}}{c}. \tag{92}$$

As for the MR strategy, active market positions are determined by comparing z_t against upper (τ_{upper}) and lower (τ_{lower}) thresholds. The trading logic is inverted to follow momentum: a **Long** position is initiated when $z_t > \tau_{\text{upper}}$, while a **Short** position is initiated when $z_t < \tau_{\text{lower}}$.

C.6 DDPM ARCHITECTURE AND TRAINING

We follow the Denoising Diffusion Probabilistic Model (DDPM) framework of Ho et al. (2020). The construction consists of a forward diffusion process that gradually adds noise to the data and a reverse process parameterized by a neural network to denoise it.

Diffusion process. We define a forward process $q(\mathbf{x}_t|\mathbf{x}_{t-1})$ which adds Gaussian noise to the data $\mathbf{x}_0$ over $T_{\text{diff}} = 1000$ timesteps according to a variance schedule β_t:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}). \tag{93}$$

We use a linear noise schedule where β_t increases linearly from $\beta_1 = 10^{-4}$ to $\beta_{T_{\text{diff}}} = 0.02$.

Network architecture. The reverse process is modelled by a 1D U-Net architecture, denoted as ϵ_θ, which takes the noisy input $\mathbf{x}_t \in \mathbb{R}^{d\times L}$, where d is the number of assets and L is the time horizon, and the timestep t to predict the added noise. The network consists of:

1. **Time Embedding:** A sinusoidal position embedding followed by a Multi-Layer Perceptron (MLP) with SiLU activation, projecting the discrete timestep t into a 256-dimensional embedding vector.
2. **Encoder:** A series of 1D residual blocks and downsampling layers. Each residual block comprises two 1D convolutional layers with Group Normalization (groups=8), SiLU activations, and a time embedding projection that is added to the intermediate feature map. Downsampling is performed via strided 1D convolutions.
3. **Bottleneck:** A sequence of residual blocks processing the latent representation at the lowest resolution.
4. **Decoder:** A symmetric path of upsampling layers (transposed 1D convolutions) and residual blocks. Skip connections concatenate feature maps from the encoder to the decoder to preserve high-frequency details.

The model is trained to minimize the mean squared error (MSE) between the true noise ϵ and the predicted noise $\epsilon_\theta(\mathbf{x}_t, t)$:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2\right], \tag{94}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. The network uses a base channel dimension of 64 and is optimized using AdamW with a learning rate of 10^{-5}.

D ADDITIONAL EMPIRICAL ANALYSES

This appendix provides supporting empirical analyses for the main findings in Section 5. The order follows the empirical evaluation: tail accuracy, checkpoint-error regimes, structural fidelity, checkpoint-selection sensitivity, and mechanism-matched aggregation controls.

D.1 TAIL-ACCURACY ROBUSTNESS

Table 4 reports tail-estimation accuracy across three tail levels, while Figures 4 and 5 show the corresponding rank-frequency behaviour of the strategy PnL distributions.

Table 4: Sensitivity of tail-risk estimation error across tail levels under the fixed budget $N = 1000$. The reported metric is average relative error of VaR and ES in percent. Lower is better.

	Synthetic			Market		
Method	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$
TailGAN	6.9 ± 3.1	5.1 ± 1.4	4.5 ± 1.2	17.8 ± 3.8	12.1 ± 1.3	10.9 ± 1.1
DDPM (EMA)	9.1 ± 3.6	7.7 ± 2.5	6.1 ± 1.9	27.3 ± 6.1	22.3 ± 4.3	19.2 ± 3.5
DDPM (DSI)	$\mathbf{4.6 \pm 0.3}$	$\mathbf{3.7 \pm 0.2}$	$\mathbf{3.4 \pm 0.1}$	$\mathbf{12.2 \pm 2.1}$	$\mathbf{9.6 \pm 1.2}$	$\mathbf{8.9 \pm 0.9}$

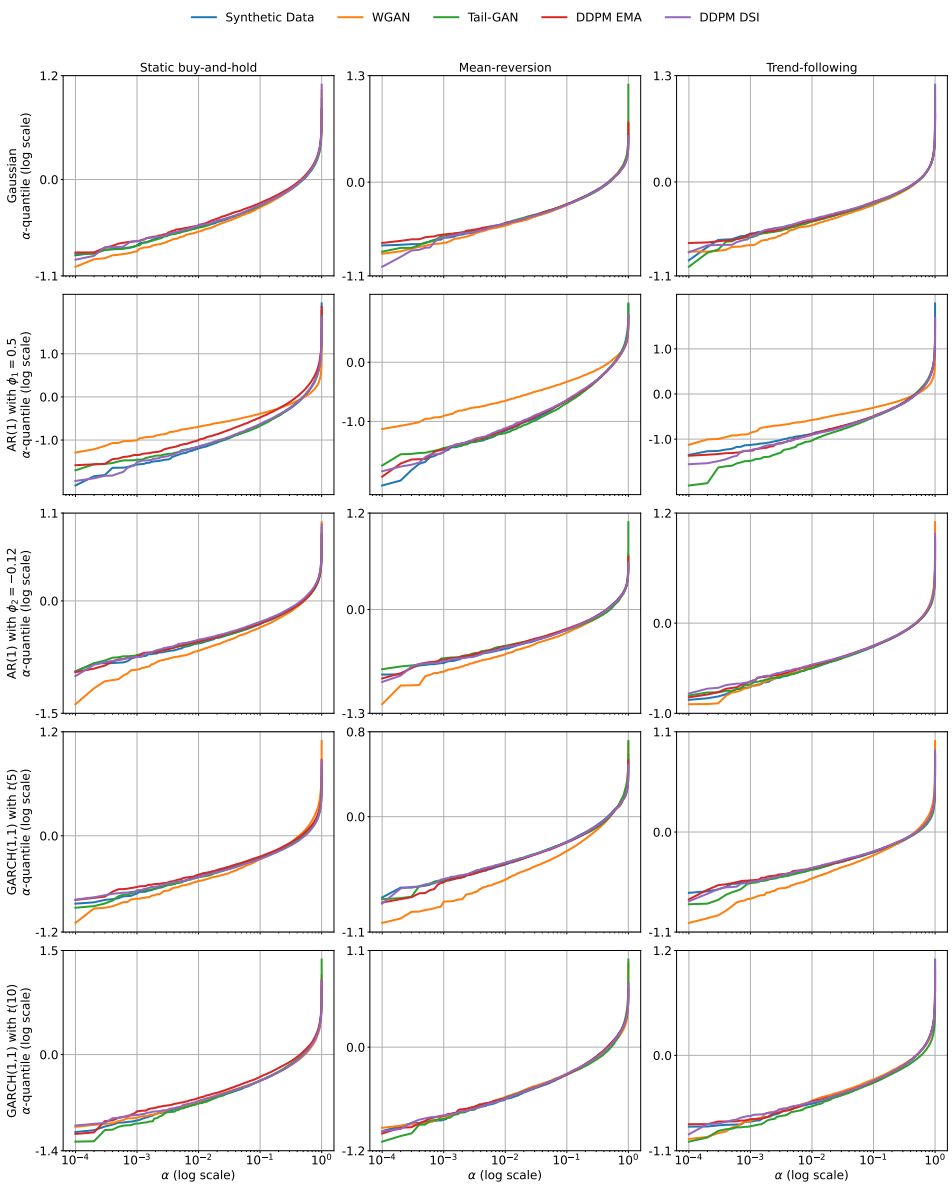


Figure 4: Tail behaviour via the synthetic rank-frequency distribution of the strategy PnL. The columns represent the strategy types.

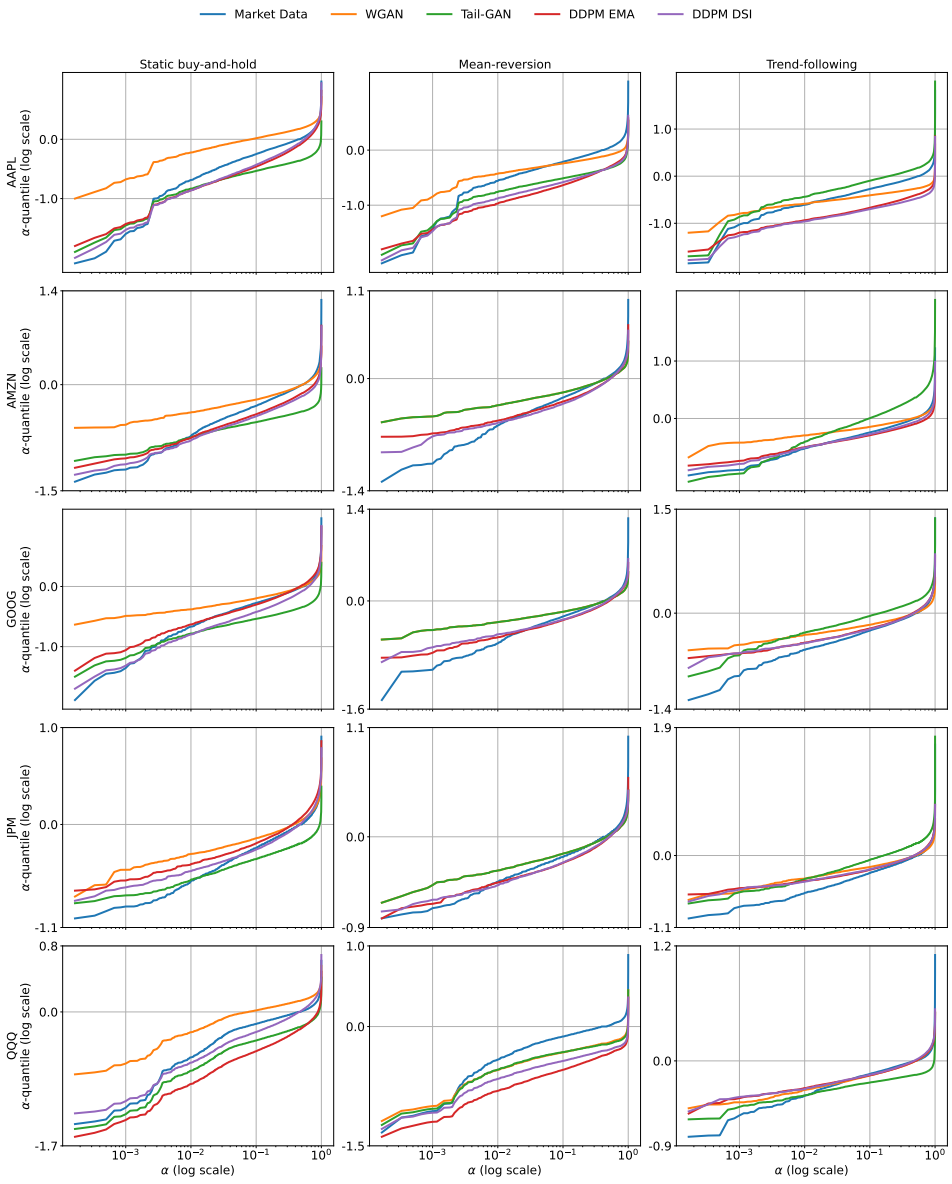


Figure 5: Tail behaviour via the market rank-frequency distribution of the strategy PnL. The columns represent the strategy types.

D.2 CHECKPOINT-ERROR REGIME DIAGNOSTICS

The following diagnostics support the regime interpretation in Sections 4 and 5. For each strategy j, let $F_{t,j}$ and $F_j^\star$ be the checkpoint and reference CDFs of the strategy-level PnL, and let $q_j^\star = q_\alpha(P_j^\star)$. We compute the integrated ES checkpoint-error process $U_{t,j}^{ES} := \int_{-\infty}^{q_j^\star}(F_{t,j}(x) - F_j^\star(x))\,dx$ at $\alpha = 5\%$, before multiplication by $1/\alpha^2$. Diagnostics are computed separately for each post-burn-in strategy–seed trajectory and then averaged over the $J = 65$ strategies and five seeds. Since $U_{t,j}^{ES}$ retains the scale of the strategy PnL, its raw squared averages are sensitive to strategy scale and isolated large errors. We therefore use them to decompose the error within each model, while Table 1 provides the cross-model comparison using strategy-normalised RE.

Table 2 in Section 5 summarizes four quantities at $K = 20$ and $M = 2$. The bias floor b_{ES}^2 measures the squared persistent component of the trajectory error. The fluctuation variance σ_{ES}^2 measures the amount of checkpoint-level variation that can potentially be averaged. The sign-consensus statistic measures the average fraction of checkpoints sharing the dominant error sign. The effective checkpoint ratio $K_{\mathrm{eff},ES}/K$ measures how much of the nominal checkpoint count remains after accounting for the strategy-specific autocorrelation of $U_{t,j}^{ES}$. At the model level, $K_{\mathrm{eff},ES}$ is variance-weighted across strategy–seed trajectories so that the aggregate trajectory decomposition is preserved. A favourable DSI regime has a small bias floor, non-negligible fluctuation variance, low sign consensus, and high $K_{\mathrm{eff},ES}/K$.

The diagnostics separate the three model trajectories. DDPM has a small bias floor and high $K_{\mathrm{eff},ES}/K$, so its checkpoint fluctuations are averageable. TailGAN has substantial fluctuation variance, but its low $K_{\mathrm{eff},ES}/K$ indicates strong autocorrelation, placing it in a correlation-limited regime. WGAN has relatively high sign consensus and a mixture-error term that remains large relative to its sampling term, indicating a shared-direction or bias-limited regime.

Table 5: Within-model finite-budget decomposition of the integrated-tail error U^{ES} at $\alpha = 5\%$, $K = 20$, $M = 2$, and $N = 1000$, averaged over $J = 65$ strategies, five diagnostic seeds, and all valid checkpoint windows. The observed MSE additionally averages 20 fixed-budget draws. Entries are raw squared integrated-tail errors before multiplication by $1/\alpha^2$ and are therefore distinct from the strategy-normalised VaR/ES RE in Table 1.

Model	Mixture-error MSE	Sampling term A_N	Predicted finite MSE	Observed fixed-budget MSE
DDPM	1.727×10^{-9}	6.807×10^{-9}	8.534×10^{-9}	8.535×10^{-9}
TailGAN	1.758×10^{-5}	1.485×10^{-8}	1.759×10^{-5}	1.760×10^{-5}
WGAN	1.340×10^{-7}	3.935×10^{-9}	1.379×10^{-7}	1.380×10^{-7}

Table 5 compares the predicted finite-budget decomposition with the observed fixed-budget MSE, averaged over strategies, seeds, all valid windows, and 20 repeated fixed-budget draws. The mixture-error MSE is the population error of the checkpoint mixture $\bar{P}_S$ relative to $P^\star$, while A_N is the stratified sampling contribution under the fixed budget. For DDPM, the sampling term dominates the reduced mixture error, so checkpoint pooling still leads to a meaningful finite-budget improvement. For TailGAN, the predicted and observed errors are dominated by the mixture-error MSE, consistent with a correlation-limited trajectory. For WGAN, the mixture-error term also dominates the sampling term, consistent with a persistent bias floor. These diagnostics explain why DSI yields large gains for DDPM in Table 1, while providing limited gains for the tested adversarial trajectories.

Figure 6 provides a complementary visual summary of checkpoint dependence across the post-burn-in trajectory. Unlike Table 2, which reports strategy-averaged diagnostics for the integrated ES process $U_{t,j}^{ES}$, this plot shows the lag-h autocorrelation of an aggregate checkpoint tail-error vector built from the strategy-level tail diagnostics used in the checkpoint analysis. The qualitative pattern is consistent with the regime interpretation above: DDPM exhibits faster decay toward zero, indicating more averageable checkpoint fluctuations, whereas TailGAN and WGAN retain stronger dependence across nearby checkpoints.

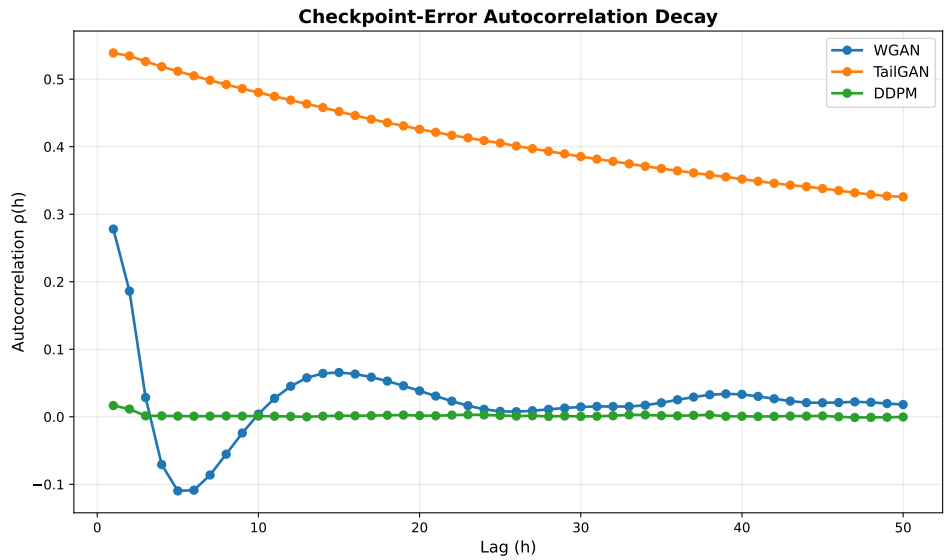


Figure 6: Autocorrelation decay of the post-burn-in checkpoint tail-error process. For each model, the curve reports the average lag-h autocorrelation across training checkpoints of a composite tail-error vector that concatenates strategy-level VaR error, ES error, and lower-tail CDF discrepancies relative to the reference distribution at $\alpha = 5\%$. Faster decay indicates that checkpoint-specific tail errors become less redundant as the stride increases, which increases the effective checkpoint count available to DSI. The DDPM curve decays most rapidly, consistent with the high K_{eff}/K values in Table 2, TailGAN remains strongly correlated, while WGAN shows more persistent dependence than DDPM.

D.3 Structural Fidelity

The following figures report cross-asset correlations and temporal autocorrelations for the synthetic and market datasets.

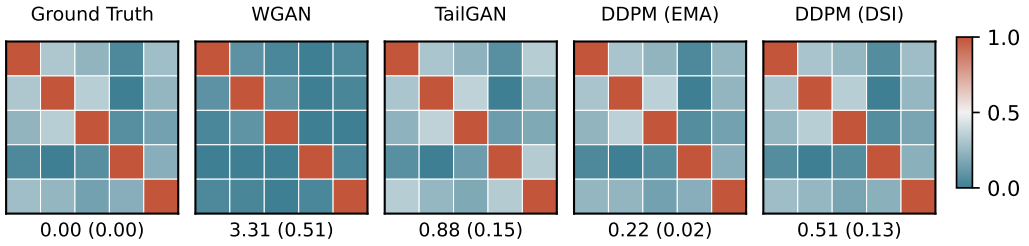


Figure 7: Multivariate correlations of the price increments in the Synthetic data and from different trained models. Numbers below each heatmap: mean and standard deviation (in parentheses) of the sum of the absolute difference between the correlation coefficients computed with all training samples and 1,000 generated samples.

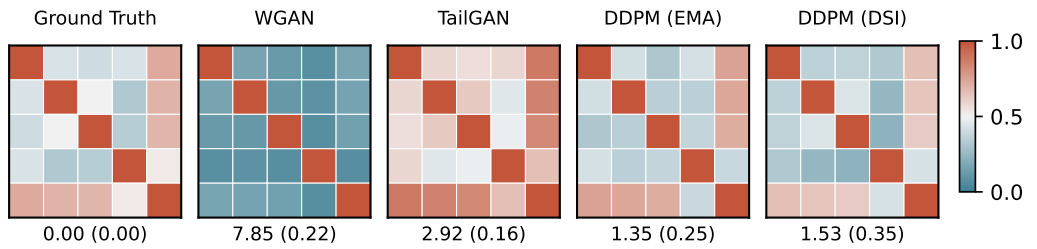


Figure 8: Cross-asset correlations of the price increments in the Market data and from different trained models. Numbers below each heatmap: mean and standard deviation (in parentheses) of the sum of the absolute difference between the correlation coefficients computed with all training samples and 1,000 generated samples.

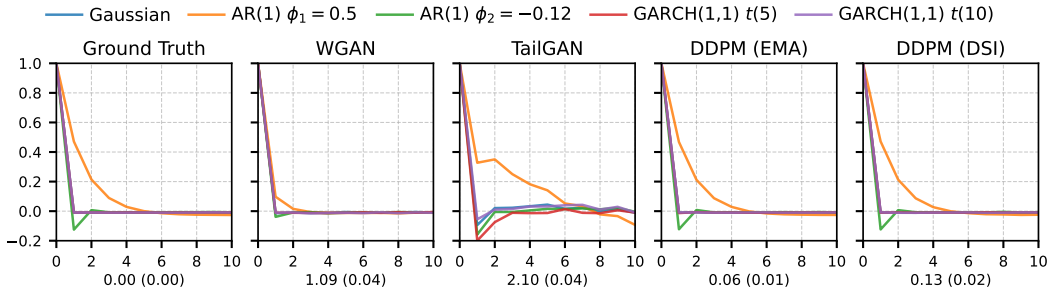


Figure 9: Auto-correlations of the price increments from different trained models for Synthetic data. Numbers below each graph: mean and standard deviation (in parentheses) of the sum of the absolute element-wise difference between auto-correlation coefficients computed with all training samples and 1,000 generated samples.

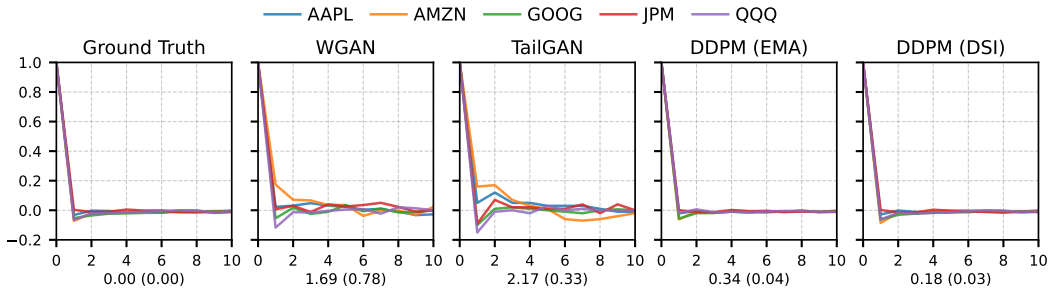


Figure 10: Auto-correlations of the price increments from different trained models for Market data. Numbers below each graph: mean and standard deviation (in parentheses) of the sum of the absolute element-wise difference between auto-correlation coefficients computed with all training samples and 1,000 generated samples.

D.4 Mixture-Size and Checkpoint-Stride Sensitivity

Stride	**Number of Checkpoints** (K)						
(M)	**5**	**10**	**20**	**50**	**100**	**150**	**200**
1	4.80 ± 0.46	4.74 ± 0.34	4.29 ± 0.18	3.84 ± 0.11	3.90 ± 0.15	4.00 ± 0.16	3.83 ± 0.17
2	4.91 ± 0.63	4.23 ± 0.41	3.63 ± 0.21	3.85 ± 0.22	3.80 ± 0.15	3.77 ± 0.12	3.94 ± 0.13
5	4.76 ± 0.62	4.47 ± 0.38	3.82 ± 0.25	3.84 ± 0.12	3.81 ± 0.19	3.85 ± 0.10	3.71 ± 0.14
10	4.71 ± 0.79	4.19 ± 0.25	4.01 ± 0.18	4.10 ± 0.27	3.66 ± 0.12	3.69 ± 0.19	3.82 ± 0.12
20	4.68 ± 0.49	4.24 ± 0.52	4.04 ± 0.31	3.95 ± 0.25	3.94 ± 0.16	4.05 ± 0.13	3.81 ± 0.12

Table 6: Full sensitivity analysis results: the reported metric is the average RE of VaR and ES in % at $\alpha = 5\%$. We trained the base DDPM for up to 5000 epochs so that we could obtain enough checkpoints for the sensitivity analysis. The validation set used for this study is independently generated from the same script defined in Appendix C.2, not the test set in the main experiments.

D.5 Mechanism-Matched Aggregation Controls

We compare DSI with EMA, SWA, SWAG, and a sample-pooled Deep Ensemble under a common sampling budget of $N = 1000$. EMA and SWA generate all trajectories from one weight-averaged model. DSI pools trajectories across K checkpoints from one training run, whereas the Deep Ensemble pools trajectories across K independently trained DDPMs. SWAG instead pools trajectories generated from parameter vectors sampled from its low-rank-plus-diagonal Gaussian weight distribution. For every mixture-based method, VaR and ES are evaluated once on the pooled sample.

Table 7: Mechanism-matched aggregation comparison at $\alpha = 0.05$ under a fixed total budget of $N = 1000$ generated trajectories. Entries are the average relative error (RE) of VaR and ES, on the same reporting scale as Table 1. Lower is better.

Method	Synthetic RE	Market RE	Training Runs
DDPM (EMA) (Ho et al., 2020)	7.7 ± 2.5	22.3 ± 4.3	1
Deep Ensemble (Lakshminarayanan et al., 2017)	3.9 ± 0.2	9.8 ± 1.3	K
SWA (Izmailov et al., 2018)	8.1 ± 2.3	19.3 ± 3.8	1
SWAG (Maddox et al., 2019)	10.9 ± 1.9	17.4 ± 4.2	1
DDPM (DSI)	$\mathbf{3.7 \pm 0.2}$	$\mathbf{9.6 \pm 1.2}$	1

Table 7 separates the effects of parameter averaging and sample-space mixing. SWA remains close to EMA, consistent with both methods reducing the training trajectory to one averaged parameter model, while SWAG is numerically worse than EMA on synthetic data but better on market data. This was expected as SWAG was used as a comparator to MC dropouts to prevent overfitting. For the Synthetic dataset the train and test sets share the same generating process so the difference between training and test distributions is much lower compared to the market dataset. The closest control is the Deep Ensemble, which achieves mean errors and dispersions comparable to DSI on both datasets. DSI therefore recovers the tail-estimation accuracy of an independently trained ensemble using checkpoints from one training trajectory rather than K independent fits. Together with the high effective checkpoint ratio in Table 2, this result indicates that spaced checkpoints provide sufficiently non-redundant tail variation to recover the benefit of independent runs.