FINHORIZON: PERSISTENT FINANCIAL DYNAMICS FROM SHORT WINDOWS TO LONG PATHS

Anonymous authors
Paper under double-blind review

ABSTRACT

The order of daily returns determines how losses cluster and drawdowns unfold. Financial scenario generation therefore requires temporal fidelity throughout long paths, even when learning from short windows. We introduce FINHORIZON, a compact neural generator that learns on 128-day windows and produces 512-day financial paths. Its key design is an explicit persistent input: stationary memory states carry latent shocks into nonlinear return dynamics, while a separate Scale Factor calibrates return magnitude after training. Across three evaluation equity indices, FINHORIZON achieves the lowest mean Dependence discrepancy among nine benchmark methods at 120, 252, and 512 days, with a 23.0% reduction relative to the strongest benchmark baseline at 512 days. Equal-length segments at the beginning, middle, and end of each rollout retain this benchmark lead, showing that the advantage extends throughout the long path. A paired intervention quantifies the contribution of persistent input: retaining memory reduces 512-day Dependence discrepancy by 31.1% relative to a white-noise input control. These results show how explicit persistence supports temporal fidelity in financial scenarios well beyond the training-window length. Code: Anonymous GitHub.

1 INTRODUCTION

Financial scenario generation turns historical returns into new market trajectories for risk analysis. The order of daily moves matters: rearranging the same returns preserves their empirical distribution and terminal return, but changes volatility clustering and maximum drawdown. A useful scenario must capture how calm and turbulent periods unfold, because this chronology determines the losses encountered along the path. These temporal regularities are central properties of financial returns (Cont, 2001; Ding et al., 1993; Corsi, 2009).

Existing research provides two complementary foundations for this task. Classical volatility models encode persistence through recursive or multiscale dynamics (Bollerslev, 1986; Corsi, 2009). Neural generators learn flexible sequence distributions using recurrent or convolutional architectures, signature-based training, and diffusion (Yoon et al., 2019; Wiese et al., 2020; Liao et al., 2024; Yuan & Qiao, 2024). These approaches motivate a focused question: how can a neural generator learn from short windows and sustain financial dependence throughout a much longer scenario? A close fit at the training-window length alone does not establish the quality of temporal structure later in a rollout.

We study unconditional single-asset generation from 128-day training windows to 512-day financial paths. Short windows provide many overlapping training examples; the required scenario spans roughly two trading years. During generation, each path evolves through latent innovations and the generator's own state. The objective is sustained financial dependence at the beginning, middle, and end of the scenario.

Temporal dependence and return amplitude pose distinct challenges (Figure 1). Shuffling can preserve the daily-return distribution while disrupting volatility clustering; positive rescaling can alter amplitude while preserving correlation-based dependence. Amplitude calibration therefore cannot restore lost temporal structure. The modeling challenge is to carry the influence of latent shocks through long rollouts while learning flexible nonlinear return dynamics, with return magnitude calibrated separately.

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.

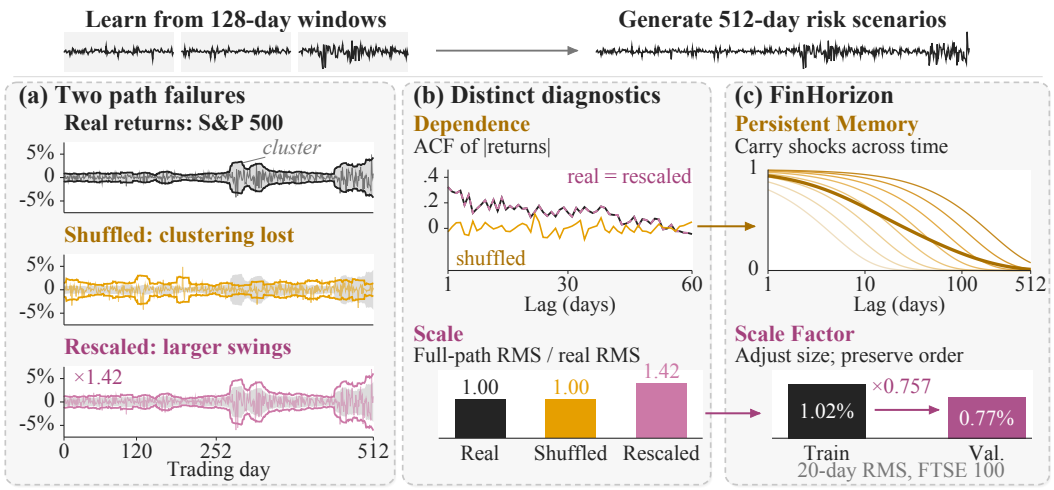


Figure 1: Long financial scenarios require both dependence and scale fidelity. (a) Shuffling S&P 500 returns disrupts clustering; rescaling enlarges daily moves. Gray envelopes show the original-path reference. (b) Absolute-return ACF detects lost dependence, while full-path RMS detects rescaling. (c) FinHorizon combines Persistent Memory with separate Scale Factor calibration. Top traces illustrate window lengths using the same real sequence; memory curves and their equal-weight mixture are illustrative, not fitted outputs. The FTSE 100 example uses pre-test 20-day RMS levels (the factor uses unrounded values).

FinHorizon addresses this challenge by making persistence an explicit input to the neural generator. Persistent Memory carries latent shocks forward through stationary decay states. The Base Generator combines this memory with current innovations to learn nonlinear return dynamics, and a separate Scale Factor calibrates amplitude after training. For the two failure modes in Figure 1, Persistent Memory supplies a temporal signal for learning how large moves cluster, while Scale Factor adjusts their magnitude. We evaluate the resulting paths through equal-length segments (position-matched) that hold evaluation length and reference data fixed, and use a paired memory intervention to measure the contribution of persistent input.

Our contributions are:

- **Explicit persistence for neural financial generation.** A compact generator couples stationary latent memory with nonlinear return dynamics. Persistence enters as a dedicated input, and separate post-training calibration sets return magnitude.
- **Temporal fidelity throughout long rollouts.** Across three evaluation equity indices, FinHorizon achieves the lowest mean Dependence discrepancy among nine benchmark methods at 120, 252, and 512 days, reducing it by 23.0% relative to the strongest benchmark baseline at 512 days. Equal-length segments at the start, middle, and end retain this benchmark lead.
- **Quantifying the contribution of persistent input.** A paired intervention shows that retaining memory reduces 512-day Dependence discrepancy by 31.1% relative to a white-noise input control.

2 Related Work

Neural path generation. TimeGAN learns recurrent latent dynamics, QuantGAN uses temporal convolutions, and SigCWGAN compares paths through signatures (Yoon et al., 2019; Wiese et al., 2020; Liao et al., 2024). COT-GAN and PCF-GAN use causal transport and characteristic functions (Xu et al., 2020; Lou et al., 2023), while Diffusion-TS and FTS-Diffusion learn diffusion generators (Yuan & Qiao, 2024; Huang et al., 2024; Takahashi & Mizuno, 2025). FinHorizon supplies a persistent memory signal to a compact neural volatility model and tests temporal fidelity throughout extended rollouts.

Persistence and scale. GARCH, HAR, and Markovian volatility approximations model persistence through recursive or multiscale states (Bollerslev, 1986; Corsi, 2009; Abi Jaber & El Euch, 2019b); structured recurrent models likewise carry information through decay modes (Gu et al., 2022; Orvieto et al., 2023). Changes in volatility level can also resemble long memory (Granger & Hyung, 2004; Mikosch & Stărică, 2004). FINHORIZON learns its memory signal jointly with the return dynamics and calibrates output magnitude after training. Matched stochastic-volatility (SV) comparisons and memory interventions test this combination against closely related volatility models. Appendix J gives the broader context.

3 METHOD

Task and modeling choices. Given chronologically split training, validation, and test returns, we learn an unconditional single-asset generator on 128-day windows and run its dynamics for up to 512 trading days. The task is to sustain financial dependence throughout the scenario while reproducing return and path-risk statistics. FINHORIZON models the stochastic component with zero conditional log-return mean. Unconditional generation means that no observed market trajectory is supplied during rollout. Fitted-mean SV controls and zero-mean refits separate persistence from accumulated drift (Section 4.5). Full task definitions are in Appendix B.

Figure 2 links the three components. Persistent Memory produces the memory signal m_t; the Base Generator combines it with current innovations to produce uncalibrated returns; the Scale Factor calibrates those returns.

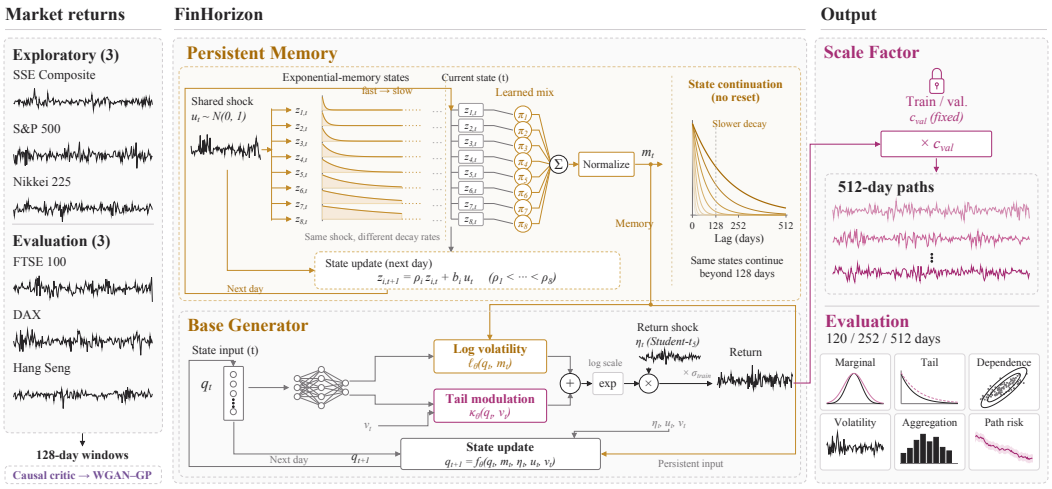


Figure 2: FINHORIZON carries persistent state from short-window learning into long-path generation. A shared shock enters eight decay states; slower states retain its influence across later days. Learned mixing and normalization produce m_t, which feeds log volatility and the recurrent-state update. The two return branches combine in log scale; both state-update loops continue without a reset at day 128. A fixed validation Scale Factor calibrates output amplitude. Three exploratory and three evaluation indices are shown; waveforms, decay responses, mixture weights, and network nodes are schematic. One generated path supplies all three evaluation prefixes.

3.1 PERSISTENT MEMORY

We carry each Gaussian shock u_t through eight fixed exponential decay states, spanning roughly one trading week to ten months. This standard Markovian construction represents persistence with a compact state (Barndorff-Nielsen, 2001; Abi Jaber & El Euch, 2019b). The model learns a positive mixture of these states:

$$z_{t+1} = R_\rho z_t + b\, u_t, \qquad m_t = \frac{\pi^\top z_t}{\sqrt{\pi^\top \Sigma_z \pi}}, \qquad \pi = \mathrm{softmax}(\vartheta). \tag{1}$$

Here R_ρ is diagonal with entries $\rho_k = \exp(-1/\tau_k)$, $b_k = \sqrt{1-\rho_k^2}$, and Σ_z is the joint stationary covariance. Initializing $z_0 \sim \mathcal{N}(0, \Sigma_z)$ supplies persistence from the first generated step; normalization gives unit readout variance for every learned mixture. The weights control the persistence profile, while a separate learned coefficient controls its influence on returns. Proposition 1 gives the stationary autocovariance; the covariance formula and interpretation are in Appendix B.3. Section 4.4 tests the roles of persistence and memory resolution separately.

3.2 Base Generator

A bounded 24-dimensional recurrent state q_t combines recent shock responses with the memory signal m_t. A learned log-volatility function s_θ modulates a centered, unit-variance Student-t_5 shock η_t, using independent Gaussian inputs ν_t:

$$\begin{aligned} q_{t+1} &= \tanh\big(Aq_t + B[\eta_t, u_t, \nu_t, |\eta_t|, m_t] + b_q\big), \\ \widetilde{Y}_{t+1} &= \sigma_{\text{train}} \exp\{s_\theta(q_t, m_t, \nu_t)\}\, \eta_t. \end{aligned} \tag{2}$$

Signed and absolute shocks allow asymmetric volatility responses; the memory signal enters both the next state and the log scale. The current return shock is independent of its scale, giving zero conditional mean. Bounded log scale and a contractive recurrent matrix control the rollout while Persistent Memory supplies the memory signal. Appendices B.2 and B.3 give the full log-scale network, dimensions, stability constraint, and fourth-moment bound.

Each memory-driving shock u_t contributes to subsequent memory states through the fixed decay dynamics. By feeding m_t into both the recurrent state and the log scale, the generator can translate this persistence into correlated return magnitudes across days. The stationary memory process supplies the same persistence structure throughout rollout, including beyond day 128; there is no reset at training-window boundaries. Training learns how this persistent input shapes nonlinear return dynamics, and long generation continues the same coupled updates with fresh innovations.

3.3 Scale Factor and fitting

With the generator frozen, we multiply every path by $c_{\text{val}} = V_{20}(r^{\text{val}})/V_{20}(r^{\text{train}})$, where V_{20} is median rolling 20-day root-mean-square volatility. This positive factor adjusts magnitude while preserving all dependence correlations and the ordering of calm and turbulent periods. It is fixed across paths and horizons and uses no test data; it is a pre-test scale calibration rather than a time-varying volatility forecast. Appendix B.4 gives the full definition.

We train jointly with WGAN–GP (Gulrajani et al., 2017) on 128-day windows, using training-only centering and standardization. Generated returns are converted back with σ_{train} without restoring the sample mean. Validation and test returns remain uncentered. Validation selects a checkpoint before scale calibration. The complete training and selection protocol is in Appendix B.5.

4 Experiments

Our experiments test FinHorizon's ability to combine sustained temporal structure with accurate long-path statistics. We measure dependence throughout extended rollouts, identify the effects of memory and calibration through paired interventions, and compare aggregate fidelity against neural, classical, and closely related SV models.

4.1 Experimental setup

Data. We use six equity indices in two groups. The SSE Composite, S&P 500, and Nikkei 225 form an exploratory group on which the model and the analysis plan, including the form of the Scale Factor, were developed. The FTSE 100, DAX, and Hang Seng form the evaluation group on which all results below are reported; the exploratory results are in Appendix I.3. Every index uses the same chronological split: 2007–2013 for training, 2014–2015 for validation, and 2016–2025 for testing. The test decade therefore starts two years after the last training return and contains several regime changes, including 2020. Returns are daily log returns of unadjusted closing levels, and Appendix B.7 gives the exact counts per interval.

Table 1: Temporal dependence at 120 days on the evaluation indices (2016–2025). Each column averages lags 1, 5, 20, and 60, the three indices, and five replications; lower is better. The $|r|$ ACF measures volatility clustering; leverage uses $\mathrm{corr}(r_t, r_{t+\ell}^2)$. Dependence is the average of the four columns. Bold and underline mark the lowest and second-lowest values in each column.

Category	Model	Return ACF ↓	$\lvert r\rvert$ ACF ↓	r^2 ACF ↓	Leverage ↓	Dependence ↓
Parametric	Student-t GJR–GARCH	0.0244	<u>0.0606</u>	0.0960	<u>0.0326</u>	0.0534
Nonparametric	Stationary bootstrap	0.0355	0.0610	0.0844	0.0491	0.0575
	Circular moving-block	0.0373	0.0735	0.0943	0.0515	0.0642
GAN	SigCWGAN	0.0327	0.0894	0.0735	0.0493	0.0612
	QuantGAN	0.0412	0.0746	<u>0.0480</u>	0.0332	<u>0.0493</u>
	TimeGAN	0.0820	0.1392	0.0789	0.0355	0.0839
Diffusion	Diffusion-TS	<u>0.0238</u>	0.0894	0.0658	0.0435	0.0556
	FTS-Diffusion	0.1492	0.1664	0.1187	0.0885	0.1307
Proposed	**FINHORIZON**	**0.0216**	**0.0558**	**0.0400**	**0.0268**	**0.0361**

Baselines. We compare against eight baselines in three groups. (1) *GAN-based generators*: SigCWGAN, TimeGAN, and QuantGAN. (2) *Diffusion-based generators*: Diffusion-TS and FTS-Diffusion. (3) *Classical models*: Student-t GJR–GARCH (Glosten et al., 1993) fitted by maximum likelihood, and the stationary and circular moving-block bootstraps (Politis & Romano, 1994; Künsch, 1989). Neural methods, including FINHORIZON, make one validation choice, a checkpoint for all of them except SigCWGAN, which selects a hyperparameter interpretation, and are run in five paired replications; classical methods have no validation choices. Configurations were frozen before their corresponding test evaluations; FTS-Diffusion and the classical comparisons followed separate protocols. Appendix B.8 records each implementation, training budget, and selection rule, and Appendix B.9 lists what was fixed in advance together with the number of fits behind the market comparison.

Training-window lengths follow the individual model protocols: FINHORIZON uses 128-day windows, whereas Diffusion-TS is fitted directly on 512-day windows and generates complete 512-day paths with the authors' model. Its shorter-horizon discrepancies use prefixes of those paths (Appendix B.8).

Evaluation. Each method generates 128 paths of 512 days per index. We score 120-, 252-, and 512-day prefixes of the same paths; 120 days is the prespecified primary horizon, supporting dependence lags up to 60 days and a larger number of real reference windows. Six metric families assess daily-return marginals, tail behavior, temporal dependence, block volatility, aggregated returns, and path risk (terminal returns and drawdowns). Dependence averages return, absolute-return, and squared-return autocorrelations and the return–future-squared-return cross-correlation at lags 1, 5, 20, and 60. All component discrepancies are standardized distances, with lower values indicating better fidelity. The fixed Overall discrepancy averages the families through four equally weighted groups. Paired 95% Student-t intervals use five replications, each averaging the same three markets; they are conditional on the observed histories. Writing the six family means as M, T, D, V, A, R, the Overall discrepancy is

$$J_h = \frac{M_h}{16} + \frac{3T_h}{16} + \frac{4D_h}{19} + \frac{3V_h}{76} + \frac{A_h}{4} + \frac{R_h}{4}. \tag{3}$$

This weighting lets improvements in temporal structure and changes in distributional fidelity be traced separately. Exact definitions are in Appendix C.1, and Appendix C.2 gives the degeneracy check and the fixed score cap that guard the aggregates.

4.2 RQ1: DOES FINHORIZON LEARN REALISTIC TEMPORAL STRUCTURE?

Evaluation protocol. For the scenario-generation task, the dependence family tests whether the model reproduces return dependence, volatility clustering, and leverage. Table 1 reports all four components at the prespecified primary horizon of 120 days.

Table 2: Temporal dependence across generation horizons. Scores use 120-, 252-, and 512-day prefixes of the same generated paths and average three evaluation markets and five replications; lower is better. Bold and underline mark the lowest and second-lowest means among all nine methods. Table 9 reports all six metric families and Overall.

Category	Model	120 days ↓	252 days ↓	**512 days ↓**
Parametric	Student-t GJR–GARCH	0.0534	0.0520	<u>0.0491</u>
Nonparametric	Stationary bootstrap	0.0575	0.0573	0.0564
	Circular moving-block	0.0642	0.0618	0.0588
GAN	SigCWGAN	0.0612	0.0612	0.0606
	QuantGAN	<u>0.0493</u>	<u>0.0496</u>	0.0508
	TimeGAN	0.0839	0.0855	0.0848
Diffusion	Diffusion-TS	0.0556	0.0572	0.0545
	FTS-Diffusion	0.1307	0.1386	0.1458
Proposed	**FINHORIZON**	**0.0361**	**0.0371**	**0.0378**

Results. FINHORIZON achieves the lowest discrepancy across all four temporal-dependence components, reducing Dependence discrepancy by approximately 26.8% relative to the strongest baseline, QuantGAN (0.0361 versus 0.0493). Paired comparisons support the improvement against all eight baselines, with every interval below zero (Appendix D.1, Table 8). Because these statistics are scale-invariant, this advantage directly demonstrates the temporal structure learned by the generator, independently of the Scale Factor.

The temporal advantage extends across markets. At 120 days, FINHORIZON outperforms all five neural baselines in mean Dependence discrepancy on each evaluation index: 0.0497 on FTSE, 0.0326 on DAX, and 0.0259 on Hang Seng. It leads all nine methods on DAX and Hang Seng; GJR–GARCH leads on FTSE (Appendix E.2).

4.3 RQ2: How well does FinHorizon preserve quality over long horizons?

Evaluation protocol. Table 2 measures Dependence on 120-, 252-, and 512-day prefixes of the same generated paths. Table 3 then compares three 120-day segments at the start, middle, and end against the same real reference windows, holding evaluation length and reference data fixed.

Results. FINHORIZON achieves the lowest mean Dependence discrepancy among the nine benchmark methods at every evaluated horizon. Its Dependence discrepancy changes from 0.0361 at 120 days to 0.0378 at 512 days, improving on the strongest 512-day baseline, GJR–GARCH (0.0491), by **23.0%**. The temporal advantage persists well beyond the 128-day training window and throughout the path: its start, middle, and end discrepancies of 0.0361, 0.0372, and 0.0382 are also the lowest benchmark means.

At 512 days, FINHORIZON also has the lowest observed mean Overall discrepancy among neural generators, 12.7% below SigCWGAN, and ranks first or second among them in every metric family (Table 9). GJR–GARCH achieves lower mean Overall discrepancy at the primary 120-day horizon (0.1840 versus 0.2519); the gap narrows to 0.0040 at 512 days (Table 10). Much of this narrowing reflects accumulated fitted drift (Appendix H).

4.4 RQ3: How do Persistent Memory and scale calibration contribute?

Evaluation protocol. Table 4 separates calibration (A) from memory (B). A toggles the Scale Factor on the same paths. B compares white-noise input (M0), single-timescale memory, and FINHORIZON's Persistent Memory. M0 replaces only the memory signal with variance-matched white noise. All retain the recurrent Base Generator and share data, seeds, training budget, checkpoint selection, and Scale Factor. Appendix E.6 tests normalization.

Table 3: Dependence discrepancy throughout 512-day paths. Equal-length segments share the same real reference data; column ranges denote trading days. Entries average three markets and five replications; lower is better. Bold marks the lowest benchmark mean. Appendix E.1 gives paired intervals and memory controls.

Model	Start (1–120) ↓	Middle (197–316) ↓	End (393–512) ↓
Student-t GJR–GARCH	0.0534	0.0478	0.0541
Stationary bootstrap	0.0575	0.0582	0.0591
Circular moving-block	0.0642	0.0615	0.0593
SigCWGAN	0.0612	0.0590	0.0608
QuantGAN	0.0493	0.0513	0.0549
TimeGAN	0.0839	0.0856	0.0851
Diffusion-TS	0.0556	0.0558	0.0543
FTS-Diffusion	0.1307	0.1606	0.1579
FINHORIZON	**0.0361**	**0.0372**	**0.0382**

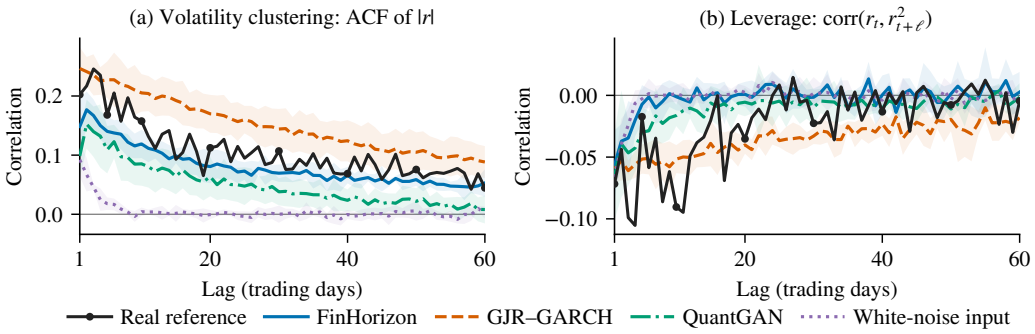


Figure 3: Persistent Memory sustains volatility clustering at the path end. Curves use days 393–512 and Table 3's real reference windows. Model curves average three markets and five replications; bands show one standard deviation across market-averaged replications. Appendix E.10 gives the reference aggregation, all market curves, and extraction protocol.

Results. Against the common white-noise input control, FINHORIZON reduces Dependence discrepancy by 32.3% at 120 days and 31.1% at 512 days; Volatility discrepancy falls by 21.9% and 24.8%, respectively. Both discrepancies also decrease at 252 days (Table 19). Figure 3 shows how this benefit appears at the path end: slowly decaying absolute-return autocorrelation is retained with Persistent Memory, while the white-noise input version approaches zero after the first few lags. Both versions retain the recurrent Base Generator; the comparison demonstrates the benefit of supplying a temporally correlated input to the learned dynamics. Together with the equal-length segment results in Table 3, this connects the memory design to sustained volatility clustering beyond the training window. FINHORIZON also reproduces negative one-day leverage in the terminal segment, although its response at longer lags is weaker than the real reference.

Persistence at the path end. On days 393–512, Persistent Memory reduces mean Dependence discrepancy by approximately 29% relative to white-noise input, from 0.0539 to 0.0382. The paired difference is -0.0158, with its interval entirely below zero (Table 13). This quantifies the benefit within a late segment, with the recurrent architecture and evaluation length held fixed. Figure 2 explains how this persistence is renewed: each new shared shock enters the decay states, and its influence reaches later returns through both log volatility and the recurrent-state update. Fresh shocks continually replenish these states throughout generation. Figure 3 resolves the resulting terminal correlation profile at every lag from 1 to 60, complementing the four lags used in the Dependence discrepancy. Together, the terminal intervention and these curves connect the continuing state dynamics to volatility clustering within the final 120 days of the generated scenario.

Table 4: Separate effects of Persistent Memory and scale calibration. A isolates scale calibration on the same paths; B compares memory inputs to the same recurrent Base Generator. Bold percentages report FINHORIZON's relative discrepancy reduction versus Scale Factor off (A) or White-noise input (B), computed from unrounded means over three markets and five replications. Lower discrepancies and higher reductions are better. Full values and paired contrasts are in Tables 17, 19, 24, and 25.

A. Calibrating return magnitude

Metric ↓	Days	Scale Factor off	**FINHORIZON**	Reduction ↑
Overall discrepancy	120	0.368	0.252	**31.6%**

B. Effect of persistent input

		Reference	With persistence		**FINHORIZON**
Metric ↓	Days	White-noise input	Single-timescale	**FINHORIZON**	Reduction ↑
Dependence	120	0.0532	0.0348	0.0361	**32.3%**
	512	0.0548	0.0374	0.0378	**31.1%**
Volatility	120	0.4678	0.3557	0.3656	**21.9%**
	512	0.4650	0.3439	0.3498	**24.8%**

Table 5: FINHORIZON's dependence advantage persists when classical baselines use all pre-test history. Classical methods fit or resample 2007–2015; FINHORIZON retains its original training and validation-based calibration. Means cover three markets and five replications; lower is better. Bold marks the lowest mean in each column. Full scores are in Table 27.

	Dependence ↓		
Model	120 days	252 days	512 days
Student-t GJR–GARCH	0.0514	0.0542	0.0558
Stationary bootstrap	0.0564	0.0544	0.0549
Circular moving-block	0.0599	0.0564	0.0555
FINHORIZON	**0.0361**	**0.0371**	**0.0378**

A single-timescale input delivers similar temporal gains (Appendix E.9). The shared ingredient is explicit persistence supplied to the recurrent generator; eight decay states provide a learnable mixture, while the intervention establishes the benefit of persistence itself.

The temporal gains come with higher Tails discrepancy, which offsets them in Overall (Appendix E.7).

Scale calibration reduces Overall discrepancy by 31.6%, from 0.368 to 0.252, while preserving Dependence discrepancy at 0.036. Its gains arise mainly on FTSE and Hang Seng, whose factors are 0.757 and 0.769; DAX changes little with a factor of 1.014.

4.5 RQ4: How robust is FinHorizon's temporal fidelity?

Evaluation protocol. We assess temporal fidelity under expanded historical information, alternative metric weights, and closely related SV specifications. Classical baselines additionally use the validation period (2007–2015 in total); SV comparisons share the market-level Scale Factor. FINHORIZON's fitting, the test split, and evaluation seeds remain unchanged. Appendices F, G, and H give the full protocols and paired comparisons.

Dependence gains with equal access to history. FINHORIZON's lower Dependence discrepancy persists at all three horizons when GJR–GARCH and both bootstraps use the complete pre-test history (Table 5). All nine paired contrasts pass Holm correction within each horizon's three classical comparisons. GJR–GARCH has lower Overall discrepancy (Table 27).

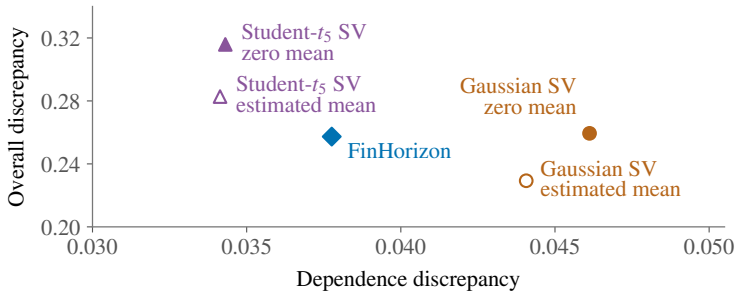


Figure 4: Temporal and overall fidelity at 512 days (lower is better). Points average three markets and five replications. All SV models use one factor and leverage, with the labeled mean specification and Student-t degrees of freedom fixed at five. All models share the market-level Scale Factor. Values and paired comparisons: Appendix H.3.

Long-path neural fidelity across metric weights. Across 10,000 six-family weight vectors drawn from a uniform Dirichlet distribution, FINHORIZON has the lowest 512-day mean discrepancy among the six neural methods in 81.61% of settings with original outputs, and 100% with common-scale outputs. These proportions characterize robustness within the neural comparison set under the sampled weights; Appendix F.1 reports the full nine-method comparison.

Dependence and overall fidelity on long paths. At 512 days, FINHORIZON combines lower mean Dependence discrepancy than both Gaussian leverage-SV controls with lower mean Overall discrepancy than both Student-t_5 controls, whose degrees of freedom are fixed at five (Figure 4). The estimated-mean Gaussian model has lower Overall discrepancy, while both Student-t_5 models have lower Dependence discrepancy. Exact values and paired comparisons are in Appendix H.3.

Temporal fidelity after drift removal. FINHORIZON retains lower mean Dependence discrepancy than GJR–GARCH at all three horizons after drift removal (post hoc subtraction of the fitted mean). This operation lowers GJR–GARCH's 512-day Overall discrepancy from 0.2533 to 0.1997. Appendix H also reports zero-mean refits, which fix the mean at zero during fitting; Appendix G examines critic channels and training objectives.

Use in long financial scenarios. FINHORIZON is suited to learning long, unconditional return scenarios from short windows when temporal fidelity is a central requirement. Volatility clustering determines whether turbulent days persist in a scenario, while leverage captures the relation between returns and subsequent volatility. Their preservation matters throughout the generated path: a realistic opening segment alone cannot establish the quality of later segments. The equal-length segment results show that FINHORIZON's benchmark dependence advantage reaches the end of 512-day rollouts. The memory intervention identifies the contribution of persistent input to this long-path temporal fidelity.

5 CONCLUSION

FINHORIZON learns long financial scenarios from short windows by coupling explicit persistence with nonlinear neural dynamics. Trained on 128-day windows, it leads the nine-method benchmark in mean Dependence discrepancy at every evaluated horizon and at the beginning, middle, and end of 512-day paths. Supplying persistent input improves both dependence and block-volatility fidelity relative to the same generator with white-noise input. These findings establish the value of explicit persistence for maintaining temporal structure throughout long neural financial scenarios.

Limitations. Our empirical findings concern unconditional single-asset generation on the three evaluated equity-index histories. The paired intervals quantify replication variability within these histories.

AI USE STATEMENT

Generative AI tools were used only to identify and summarize literature and improve manuscript readability. They were not used for theory, hypotheses, methodology, experiments, implementation, translation, data processing, analysis, or result interpretation. The authors verified every cited source and all AI-assisted text and take full responsibility.

ETHICS STATEMENT

This study uses publicly available historical equity-index prices and model-generated return paths. It involves no human participants, personal data, or real trades. The generated scenarios are intended for research on financial time-series modeling and do not constitute investment advice.

REPRODUCIBILITY STATEMENT

Section 4 describes the chronological data splits, baseline comparisons, evaluation protocol, and paired replications. Appendix B provides the model architecture, training and calibration procedures, data sources, baseline configurations, and random seeds. Appendix C defines the evaluation metrics and aggregation rules. An anonymous implementation is linked in the abstract.

REFERENCES

Eduardo Abi Jaber and Omar El Euch. Markovian structure of the Volterra Heston model. *Statistics & Probability Letters*, 149:63–72, 2019a. doi: 10.1016/j.spl.2019.01.024.

Eduardo Abi Jaber and Omar El Euch. Multifactor approximation of rough volatility models. *SIAM Journal on Financial Mathematics*, 10(2):309–349, 2019b. doi: 10.1137/18M1170236.

Yihao Ang, Qiang Huang, Yifan Bao, Anthony K. H. Tung, and Zhiyong Huang. TSGBench: Time series generation benchmark. *Proceedings of the VLDB Endowment*, 17(3):305–318, 2023. doi: 10.14778/3632093.3632097. URL https://doi.org/10.14778/3632093.3632097.

Richard T. Baillie, Tim Bollerslev, and Hans Ole Mikkelsen. Fractionally integrated generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 74(1):3–30, 1996. doi: 10.1016/S0304-4076(95)01749-6. URL https://doi.org/10.1016/S0304-4076(95)01749-6.

Ole E. Barndorff-Nielsen. Superposition of Ornstein–Uhlenbeck type processes. *Theory of Probability and Its Applications*, 45(2):175–194, 2001. doi: 10.1137/S0040585X97978166.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986. doi: 10.1016/0304-4076(86)90063-1.

Damien Challet and Vincent Ragel. Multi-timescale recurrent neural networks beat rough volatility for intraday volatility prediction. *Risks*, 12(6):84, 2024. doi: 10.3390/risks12060084. URL https://doi.org/10.3390/risks12060084.

Xiaohong Chen and Yanqin Fan. Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, 130(2):307–335, 2006. doi: 10.1016/j.jeconom.2005.03.004. URL https://doi.org/10.1016/j.jeconom.2005.03.004.

Rama Cont. Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001. doi: 10.1080/713665670.

Fulvio Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009. doi: 10.1093/jjfinec/nbp001.

Zhuanxin Ding, Clive W. J. Granger, and Robert F. Engle. A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, 1(1):83–106, 1993. doi: 10.1016/0927-5398(93)90006-D. URL https://doi.org/10.1016/0927-5398(93)90006-D.

Jim Gatheral, Thibault Jaisson, and Mathieu Rosenbaum. Volatility is rough. *Quantitative Finance*, 18(6):933–949, 2018. doi: 10.1080/14697688.2017.1393551.

Lawrence R. Glosten, Ravi Jagannathan, and David E. Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5): 1779–1801, 1993. doi: 10.1111/j.1540-6261.1993.tb05128.x.

C. W. J. Granger. Long memory relationships and the aggregation of dynamic models. *Journal of Econometrics*, 14(2):227–238, 1980. doi: 10.1016/0304-4076(80)90092-5. URL https://doi.org/10.1016/0304-4076(80)90092-5.

Clive W. J. Granger and Namwon Hyung. Occasional structural breaks and long memory with an application to the S&P 500 absolute stock returns. *Journal of Empirical Finance*, 11(3): 399–421, 2004. doi: 10.1016/j.jempfin.2003.03.001. URL https://doi.org/10.1016/j.jempfin.2003.03.001.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5767–5777, 2017.

Hongbin Huang, Minghua Chen, and Xiao Qiao. Generative learning for financial time series with irregular and scale-invariant patterns. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CdjnzWsQax.

Daniel Jarrett, Ioana Bica, and Mihaela van der Schaar. Time-series generation by contrastive imitation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 28968–28982. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f2b4053221961416d47d497814a8064f-Abstract.html.

Paul Jeha, Michael Bohlke-Schneider, Pedro Mercado, Shubham Kapoor, Rajbir Singh Nirwan, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. PSA-GAN: Progressive self attention GANs for synthetic time series. In *The Tenth International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Ix_mh42xq5w.

Karen A. Kopecky and Richard M. H. Suen. Finite state markov-chain approximations to highly persistent processes. *Review of Economic Dynamics*, 13(3):701–714, 2010. doi: 10.1016/j.red.2010.02.002.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2241719.

Shujian Liao, Hao Ni, Marc Sabaté-Vidales, Łukasz Szpruch, Magnus Wiese, and Baoren Xiao. Sig-Wasserstein GANs for conditional time series generation. *Mathematical Finance*, 34(2): 622–670, 2024. doi: 10.1111/mafi.12423.

Hang Lou, Siran Li, and Hao Ni. PCF-GAN: Generating sequential data via the characteristic function of measures on the path space. In *Advances in Neural Information Processing Systems*, volume 36, pp. 39755–39781, 2023. doi: 10.52202/075280-1728.

Thomas Mikosch and Cătălin Stărică. Nonstationarities in financial time series, the long-range dependence, and the IGARCH effects. *The Review of Economics and Statistics*, 86(1):378–390, 2004. doi: 10.1162/003465304323023886. URL https://doi.org/10.1162/003465304323023886.

John Miller and Moritz Hardt. Stable recurrent models. In *The Seventh International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hygxb2CqKm.

Ilan Naiman, Nimrod Berman, Itai Pemper, Idan Arbiv, Gal Fadlon, and Omri Azencot. Utilizing image transforms and diffusion models for generative modeling of short and long time series. In *Advances in Neural Information Processing Systems*, volume 37, pp. 121699–121730. Curran Associates, Inc., 2024. doi: 10.52202/079017-3868.

Antonio Orvieto, Samuel L. Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26670–26698. PMLR, 2023.

Dimitris N Politis and Joseph P Romano. A circular block-resampling procedure for stationary data. Purdue University. Department of Statistics, 1991.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994. doi: 10.1080/01621459.1994.10476870.

Tomonori Takahashi and Takayuki Mizuno. Generation of synthetic financial time series by diffusion models. *Quantitative Finance*, 25(10):1507–1516, 2025. doi: 10.1080/14697688.2025.2528697.

Yuki Tanaka, Ryuji Hashimoto, Takehiro Takayanagi, Zhe Piao, Yuri Murayama, and Kiyoshi Izumi. CoFinDiff: Controllable financial diffusion model for time series generation. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 9357–9365, 2025. doi: 10.24963/ijcai.2025/1040. URL https://doi.org/10.24963/ijcai.2025/1040.

Magnus Wiese, Robert Knobloch, Ralf Korn, and Peter Kretschmer. Quant GANs: Deep generation of financial time series. *Quantitative Finance*, 20(9):1419–1440, 2020. doi: 10.1080/14697688.2020.1730426.

Tianlin Xu, Li K. Wenliang, Michael Munn, and Beatrice Acciaio. COT-GAN: Generating sequential data via causal optimal transport. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 5508–5518, 2019.

Xinyu Yuan and Yan Qiao. Diffusion-TS: Interpretable diffusion for general time series generation. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=4h1apFjO99.

Linqi Zhou, Michael Poli, Winnie Xu, Stefano Massaroli, and Stefano Ermon. Deep latent state space models for time-series generation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42625–42643. PMLR, 2023.

APPENDIX GUIDE

The appendices follow the evidence used in the main paper.

Question	Main paper	Evidence
How is the model specified and evaluated?	Sections 3–4.1	Appendices A–C
How do benchmark methods compare?	Tables 1–2	Appendix D
Does dependence persist, and what does memory contribute?	Tables 3–4, Figure 3	Appendix E
Do history, weights, SV, or drift explain the results?	Section 4.5, Figure 4	Appendices F–H
What do process and calibration checks show?	—	Appendix I
How does the design relate to earlier work?	Section 2	Appendix J

Analysis sequence. The primary evaluation and later extensions used the same held-out histories. The table distinguishes rules fixed before the original evaluation from additional analyses designed with earlier results known.

Analysis	Specification sequence
Primary evaluation	Markets, splits, seeds, calibration, primary horizon, metrics, and interval analysis fixed before evaluation-index scoring; Appendix B.9 details separately frozen baseline extensions.
Trajectory and memory	Designed after original results; positions, lags, control, and reporting rules fixed before the extension. Original samples and weights retained (Appendix E).
Weights, history, SV	Designed after primary results; new fitting/scoring rules fixed before new comparisons. Common-scale SV analysis followed unscaled SV results (Appendix F).
Critic, drift, leverage	Sequential extensions fixed before their own results; leverage implementation amended before any leverage fit. Fixed-zero training-objective comparison followed the drift diagnostic (Appendix G).
Mean–scale controls	Specified after earlier SV results and before new GJR–GARCH scores and zero-mean leverage fits; contrast families recorded during analysis (Appendix H).

A Mathematical Details

The following proposition characterizes the stationary memory signal.

Proposition 1 (Stationarity and autocovariance of the memory)**.** *Let $z_0 \sim \mathcal{N}(0, \Sigma_z)$ be independent of the IID standard Gaussian sequence $(u_t)_{t\geq 0}$, and let m_t be given by Equation (1). Then (z_t) is strictly stationary, and for every integer $\ell \geq 0$,*

$$\mathrm{Cov}(m_t, m_{t+\ell}) = \sum_{j=1}^{8} \beta_j \rho_j^{\ell}, \qquad \beta_j > 0, \qquad \sum_{j=1}^{8} \beta_j = 1. \tag{4}$$

A.1 Proof of Proposition 1

Let $w = \pi/(\pi^\top \Sigma_z \pi)^{1/2}$, so that $m_t = w^\top z_t$ and $w^\top \Sigma_z w = 1$. Suppose $z_t \sim \mathcal{N}(0, \Sigma_z)$ and u_t is independent of z_t. The linear update gives a Gaussian next state with mean zero and covariance

$$\mathrm{Cov}(z_{t+1}) = R_\rho \Sigma_z R_\rho^\top + bb^\top = \Sigma_z,$$

where the last equality follows entrywise from $(\Sigma_z)_{ij} = b_i b_j/(1-\rho_i\rho_j)$. The Gaussian law $\mathcal{N}(0, \Sigma_z)$ is therefore invariant; initializing in that law and using a time-homogeneous transition gives strict stationarity. For $\ell \geq 0$, $\mathrm{Cov}(z_t, z_{t+\ell}) = \Sigma_z R_\rho^\ell$, hence

$$\mathrm{Cov}(m_t, m_{t+\ell}) = w^\top \Sigma_z R_\rho^\ell w = \sum_{j=1}^{8} w_j (\Sigma_z w)_j \, \rho_j^{\ell}.$$

Every entry of w and Σ_z is positive, so $\beta_j = w_j(\Sigma_z w)_j > 0$, and $\sum_j \beta_j = w^\top \Sigma_z w = 1$. □

A.2 Fourth-moment bound

Write $s_\theta = \ell_t + \kappa_t$ with $\ell_t = 1.5\tanh(h_\sigma(q_t) + \lambda m_t)$ and $\kappa_t = 0.75\tanh(h_\kappa([q_t, \nu_t]))$, so $|\ell_t + \kappa_t| \leq 2.25$. A unit-variance Student-t_5 variable has fourth moment 9. Therefore

$$\mathbb{E}|\bar{Y}_{t+1}|^4 = \mathbb{E}\big[\exp\{4(\ell_t + \kappa_t)\}\,|\eta_t|^4\big] \leq 9e^9 < \infty.$$

Since $\widetilde{Y}_{t+1} = \sigma_{\mathrm{train}} \bar{Y}_{t+1}$ and $Y_{t+1} = c_{\mathrm{val}} \widetilde{Y}_{t+1}$ with finite positive constants, $\mathbb{E}|Y_{t+1}|^4 \leq 9c_{\mathrm{val}}^4 \sigma_{\mathrm{train}}^4 e^9 < \infty$. No independence between the conditional log scale and η_t is needed.

A.3 FINITE-HORIZON EXPONENTIAL APPROXIMATION

The following proposition states the capacity of a finite positive exponential basis to approximate power-law decay over a bounded range of lags; it concerns the basis, not the learned weights.

Proposition 2 (Finite-horizon exponential approximation)**.** *Let $k(t) = (t+\delta)^{-\alpha}$ with $\alpha, \delta > 0$, fix $T_{\max} < \infty$, and choose $\Delta_x > 0$, $x_0 \in \mathbb{R}$, and $K \geq 1$. Define*

$$s_j = e^{x_j}, \qquad \omega_j = \frac{\Delta_x}{\Gamma(\alpha)} \exp\big(\alpha x_j - \delta e^{x_j}\big), \qquad x_j = x_0 + j\Delta_x, \quad j = 0, \dots, K-1, \tag{5}$$

and $\widetilde{k}_K(t) = \sum_{j=0}^{K-1} \omega_j e^{-s_j t}$. All ω_j and s_j are positive. If $\delta e^{x_0 + K\Delta_x} \geq \alpha$, then

$$\sup_{0 \leq t \leq T_{\max}} \left| \frac{\widetilde{k}_K(t)}{k(t)} - 1 \right| \leq E_{\mathrm{disc}}(\Delta_x) + \frac{(T_{\max}+\delta)^\alpha}{\Gamma(\alpha)}(R_- + R_+), \tag{6}$$

$$E_{\mathrm{disc}}(\Delta_x) = \frac{2}{\Gamma(\alpha)} \sum_{m=1}^{\infty} \left| \Gamma\Big(\alpha + \frac{2\pi i m}{\Delta_x}\Big) \right|,$$

$$R_- = \frac{\Delta_x e^{\alpha(x_0 - \Delta_x)}}{1 - e^{-\alpha\Delta_x}},$$

$$R_+ = \Delta_x e^{\alpha x_+ - \delta e^{x_+}} + \delta^{-\alpha}\Gamma(\alpha, \delta e^{x_+}), \qquad x_+ = x_0 + K\Delta_x,$$

where $\Gamma(\alpha, \cdot)$ is the upper incomplete gamma function.

Proof. The gamma integral and the change of variables $s = e^x$ give

$$k(t) = \frac{1}{\Gamma(\alpha)} \int_0^\infty s^{\alpha-1} e^{-(t+\delta)s}\, ds = \frac{1}{\Gamma(\alpha)} \int_{-\infty}^\infty e^{\alpha x - (t+\delta)e^x}\, dx.$$

Applying the infinite trapezoidal rule on the lattice $x_0 + j\Delta_x$ and Poisson summation, the nonzero Fourier modes have magnitudes $|\Gamma(\alpha + 2\pi i m/\Delta_x)|/\Gamma(\alpha)$ relative to $k(t)$, with unit-modulus phases; summing positive and negative frequencies gives $E_{\mathrm{disc}}(\Delta_x)$. Truncating the lattice to $j = 0, \dots, K-1$ omits lower and upper tails. For the lower tail, dropping $e^{-(t+\delta)e^x} \leq 1$ leaves the geometric sum R_-. For the upper tail, $g(x) = e^{\alpha x - \delta e^x}$ is decreasing once $\delta e^x \geq \alpha$, which holds at x_+ by assumption, and the integral test gives $\Delta_x \sum_{j \geq K} g(x_0 + j\Delta_x) \leq \Delta_x g(x_+) + \int_{x_+}^\infty g = R_+$. Finally $k(t)^{-1} = (t+\delta)^\alpha \leq (T_{\max}+\delta)^\alpha$ on $[0, T_{\max}]$. Combining the three terms proves Equation (6); positivity follows from Equation (5). □

For $\alpha = 0.44$, a constructive logarithmic grid has maximum relative error 3.69% at $T_{\max} = 512$ with 32 modes (bound 3.79%) and 0.179% with 48 modes (bound 0.179%). Figure 5 reports these certificates together with positive covariance mixtures fitted on the main narrow eight-rate bank and on the auxiliary wide bank of Appendix B.6.

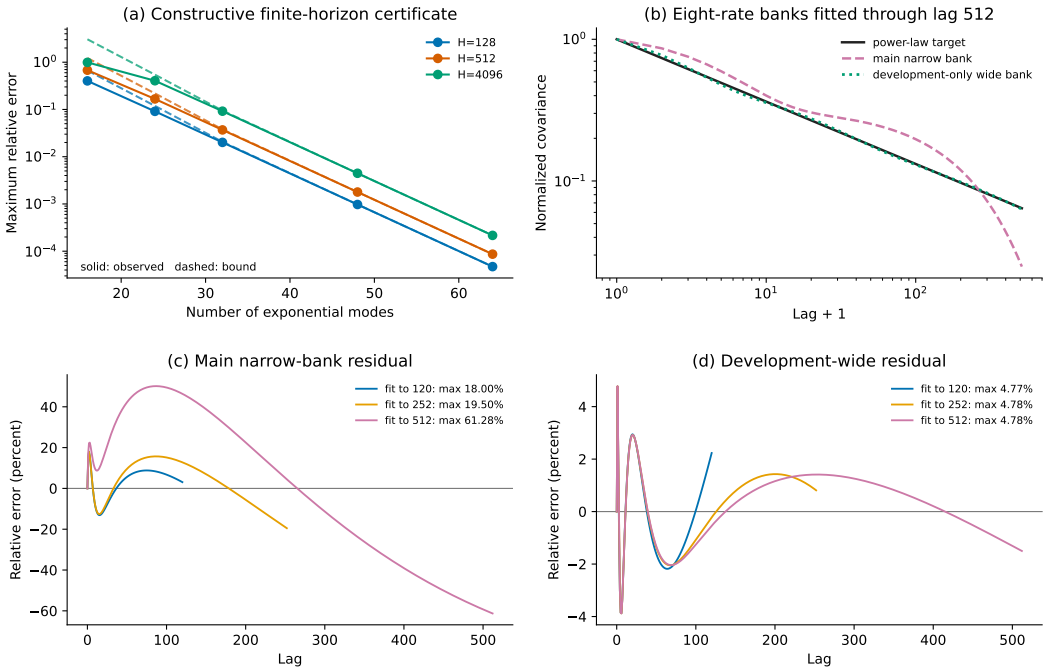


Figure 5: Finite-horizon approximation of $k(t) = (t+1)^{-0.44}$. (a) Observed maximum errors of constructive logarithmic grids and the corresponding certificates from Proposition 2. (b) Positive covariance mixtures on the main narrow bank and the auxiliary wide bank, each fitted through lag 512. (c) Residuals of the main narrow bank; maximum errors through lags 120, 252, and 512 are 18.0%, 19.5%, and 61.3%. (d) Residuals of the auxiliary wide bank; its lag-512 fit has 4.78% maximum and 1.21% RMS relative error. The market model uses the narrow bank.

B Model and Training Details

B.1 Task, mean specification, and return units

The generator maps independent random inputs $\xi_{1:h}$ to an unconditional return path $Y_{1:h} = G_\theta(\xi_{1:h})$. It learns from 128-day windows of the training history and is evaluated at 120, 252, and 512 days; window length is distinct from the total available history. Training returns are centered by their training mean and divided by the population standard deviation σ_{train}. Generated returns are multiplied by σ_{train} without restoring the sample mean. The model therefore specifies a zero-mean stochastic component, not an estimate of the market's expected return. Validation and test returns remain uncentered. Validation selects the checkpoint and then calibrates scale; test returns enter scoring only. Section 4.5 compares fitted-mean controls and zero-mean refits.

B.2 Complete generator specification

The nonlinear state is $q_t \in \mathbb{R}^{24}$ and the memory state is $z_t \in \mathbb{R}^8$. Across time, the triples (η_t, u_t, ν_t) are IID with mutually independent coordinates:

$$\eta_t = \sqrt{3/5}\, T_t, \quad T_t \sim t_5, \qquad u_t \sim \mathcal{N}(0,1), \qquad \nu_t \sim \mathcal{N}(0, I_6),$$

so η_t has mean zero and variance one. The stationary covariance of the memory has the closed form $(\Sigma_z)_{ij} = b_i b_j/(1-\rho_i\rho_j)$ and satisfies $\Sigma_z = R_\rho \Sigma_z R_\rho^\top + bb^\top$. Distinct decay rates give linearly independent impulse responses, hence a positive-definite Σ_z and a valid Cholesky factor L_z; we initialize $z_0 = L_z e_0$ with $e_0 \sim \mathcal{N}(0, I_8)$. The decay rates and L_z are fixed. The trainable memory parameters are the mixture logits $\vartheta \in \mathbb{R}^8$, initialized to zero so that π starts uniform, and the direct coefficient $\lambda = \text{sigmoid}(a) \in (0,1)$, initialized at $\lambda = 0.10$.

For $\psi_t = [\eta_t, u_t, \nu_t^\top, |\eta_t|, m_t]^\top \in \mathbb{R}^{10}$, the exact update and return equations are

$$q_{t+1} = \tanh(Aq_t + B\psi_t + b_q), \qquad \|A\|_2 \le 0.8,$$
$$\ell_t = 1.5\tanh\big(h_\sigma(q_t) + \lambda m_t\big), \qquad \kappa_t = 0.75\tanh\big(h_\kappa([q_t, \nu_t])\big),$$
$$\bar{Y}_{t+1} = \exp(\ell_t + \kappa_t)\,\eta_t, \qquad \widetilde{Y}_{t+1} = \sigma_{\text{train}}\bar{Y}_{t+1},$$

so $s_\theta = \ell_t + \kappa_t$ in Equation (2). Here $h_\sigma : \mathbb{R}^{24} \to \mathbb{R}$ is affine and h_κ is a two-layer network with input dimension 30, one 72-unit $\tanh$ hidden layer, and scalar output. The recurrent matrix is initialized orthogonally before its norm bound is enforced; h_σ and the final layer of h_κ start at zero. The last column of B carries m_t into the recurrent state, and λ gives its separate direct contribution to ℓ_t; the two coefficients measure different influences. We set $q_0 = 0$ and discard 32 generated steps before retaining a path. For two nonlinear states driven by the same inputs, $\|q_t - q'_t\|_2 \le 0.8^t\|q_0 - q'_0\|_2$; this is conditional forgetting of the initialization, not a mixing claim for the full return process.

Component	Parameters	Calculation
Memory mixture logits	8	K
Base-generator RNN	864	$24 \cdot 10 + 24 \cdot 24 + 2 \cdot 24$
Log-scale head	25	$24 + 1$
Bounded memory loading	1	scalar logit
Tail MLP	2,305	$(30 \cdot 72 + 72) + (72 + 1)$
Total	3,203	

B.3 Memory normalization and generator properties

All memory states share the same innovation and are correlated. Consequently, the variance of $\pi^\top z_t$ changes with the mixture. The denominator $\sqrt{\pi^\top \Sigma_z \pi}$ in Equation (1) fixes the readout variance at one for every π: mixture weights control the persistence profile, while λ and the memory-input column of B control its influence. This is a parameterization guarantee; Appendix E.6 tests its empirical effect. The fixed covariance also allows stationary initialization, avoiding a long transient in the slow states. Proposition 1 gives the stationary autocovariance, and Appendix A.3 quantifies finite-horizon exponential approximation for the chosen grids.

The current return innovation is centered and independent of the states and Gaussian input setting its scale, so its conditional mean is zero. Signed and absolute innovations enter the next recurrent state, allowing asymmetric responses. The bounded log scale lies in $[-2.25, 2.25]$, which gives the fourth-moment bound in Appendix A.2. The recurrent constraint controls dependence on initial state; persistent input can continue to influence both the state and the log scale.

B.4 Scale calibration and its invariances

For a return sequence x, define

$$V_{20}(x) = \text{median}_{t \ge 20}\left(\frac{1}{20}\sum_{j=0}^{19} x_{t-j}^2\right)^{1/2}, \qquad c_{\text{val}} = \frac{V_{20}(r^{\text{val}})}{V_{20}(r^{\text{train}})}, \qquad Y_{1:h} = c_{\text{val}}\widetilde{Y}_{1:h}. \quad (7)$$

The median summarizes a trading month's rolling volatility robustly. The ratio is computed once after checkpoint selection and is fixed across paths, horizons, and replications of a market. No generated or test return enters its calculation. For the correlation statistics Φ_{dep} defining Dependence,

$$\Phi_{\text{dep}}(c\,y) = \Phi_{\text{dep}}(y), \qquad c > 0. \quad (8)$$

Positive scaling preserves these correlations and the ordering of calm and turbulent periods, while changing marginal distributions, tails, block volatility, aggregated returns, and path risk. It calibrates magnitude without creating volatility clustering and is fixed before the test period, rather than forecasting its time-varying volatility.

B.5 Optimization and selection

The critic is a causal one-dimensional convolutional network with four input channels $(r_t, |r_t|, \log(1 + r_t^2), \sum_{s \le t} r_s)$, 32 channels per layer, and dilations $1, 2, \ldots, 256$; it has 28,515 parameters. We

use WGAN–GP with penalty coefficient 10, five critic updates per generator update, batch size 24, and 1,000 generator updates. Both optimizers are Adam with learning rate 2×10^{-4} and $(\beta_1, \beta_2) = (0.5, 0.9)$. Generator gradients are clipped to norm 10, and the recurrent operator norm is projected below 0.8. Training uses non-wrapping 128-day windows from the standardized training interval. Candidate checkpoints occur after 250, 500, and 1,000 generator updates; we retain the earliest candidate whose 120-day validation score is within one percent of the minimum, with all candidates scored on the same validation windows and random draws. The selected generator draws one length-512 sample per replication, and the 120- and 252-day evaluations use its prefixes.

Numerical projection in the simulated grid audit. The repeated simulated grid used target $c(1 - 16d\epsilon_{32})$ with $c = 0.8$, $d = 24$, and float32 precision ϵ_{32}, while retaining the mathematical cap c. This margin addresses round-off at the projection boundary.

B.6 Decay grids

The main narrow grid uses eight geometrically spaced timescales between $-1/\log(0.8)$ and $-1/\log(0.995)$ trading days:

$$(4.4814,\ 7.7076,\ 13.2563,\ 22.7996,\ 39.2131,\ 67.4426,\ 115.9947,\ 199.4996).$$

The auxiliary wide bank used in ablations instead uses $(2, 6, 18, 54, 162, 486, 1458, 4374)$. It has the same eight filter coordinates, 24 fast coordinates, emission heads, and 3,203 trainable parameters; under a common seed all trainable tensors are initialized identically, and only the fixed dynamics derived from the grid differ.

B.7 Data splits

Table 6 gives the exact number of daily log returns in each interval. The split is by calendar date, so the counts differ slightly across indices because of national holidays.

Table 6: Equity-index samples and chronological design. Counts are daily log returns in the common training (2007–2013), validation (2014–2015), and test (2016–2025) periods. The exploratory group informed the model and the analysis plan, including the form of the Scale Factor; all reported comparisons use the evaluation group.

Role	Equity index	N_{train}	N_{val}	N_{test}
Exploratory	SSE Composite	1,699	489	2,430
	S&P 500	1,762	504	2,514
	Nikkei 225	1,713	488	2,444
Evaluation	FTSE 100	1,768	506	2,525
	DAX	1,777	505	2,537
	Hang Seng	1,725	494	2,459

The SSE Composite is obtained from BaoStock and the remaining indices from Yahoo Finance.[1] Consecutive unadjusted closing levels are converted to daily decimal log returns. No model is re-estimated inside the test period.

B.8 Baseline configurations

Table 7 records the implementation, training budget, and selection rule of every method in the market comparison. Candidate schedules are set separately per method because one generator update means different things across architectures. For methods with checkpoint candidates we keep the earliest candidate whose 120-day validation score is within one percent of that method's best, which uses no test data. SigCWGAN has one trained model per hyperparameter interpretation and uses the global rule described in the table. The three classical methods have one training-only fit each and no validation selection at all.

[1] Data sources: `http://baostock.com` and `https://finance.yahoo.com`.

Table 7: Implementations and selection rules in the market comparison.

Model	Implementation	Training and selection rule
Student-t GJR–GARCH	Maximum likelihood with Student-t innovations	Three fixed starts, at most 1,000 L-BFGS-B iterations per start, minimum training negative log likelihood, 2,000-step sampling burn-in; no validation selection
Stationary bootstrap	Training-return resampling with geometric block lengths	Expected block length 64; no learned parameter, checkpoint, or validation selection
Circular moving-block	Circular block resampling of training returns	Fixed block length 64; no learned parameter, checkpoint, or validation selection
FINHORIZON	Our implementation (Appendix B.2)	1,000 generator updates; candidates at 250, 500, and 1,000
SigCWGAN	Our implementation of the conditional expected-signature objective. Official code for reference: `https://github.com/SigCGANs/Conditional-Sig-Wasserstein-GANs`	One trained model per interpretation after 1,000 generator updates; within each replication one interpretation is selected for all three indices by the equal-index validation mean, retaining paper-primary when it is within 1% of the minimum
TimeGAN	PyTorch adaptation of the official code: `https://github.com/jsyoon0823/TimeGAN`	1,000 embedder, 1,000 supervisor, and 1,000 joint updates; joint-stage candidates at 100, 500, and 1,000
QuantGAN	Our temporal-convolutional implementation	10,000 generator updates; candidates at 1,000, 3,000, and 10,000
Diffusion-TS	Official code: `https://github.com/Y-debug-sys/Diffusion-TS`. 512-day training and generation windows, with 120/252-day prefix evaluation	10,000 training updates; candidates at 1,000, 3,000, and 10,000; 500 diffusion and 50 sampling steps
FTS-Diffusion	Our implementation with a modified signed-magnitude sampler. Official supplementary code available from: `https://proceedings.iclr.cc/paper_files/paper/2024/hash/f90fc76b199fe6b0ec2a51aaf72c3277-Abstract-Conference.html`	Two separately sealed market cohorts of 15 fits each: evaluation and exploratory indices; candidates at 100, 200, and 400; earliest checkpoint within 1% of the best 120-day validation score; standard raw output with no scale factor

The FTS-Diffusion entry is not a reproduction of author code, and no upstream implementation was imported. Its sampler amendment passes every finite signed magnitude unchanged into reconstruction and rejects only a nonfinite magnitude or a nonfinite or nonpositive reconstructed price path, applying no clipping, absolute value, softplus transformation, or scale factor. Each three-index cohort contains 15 fits whose selected candidates and checkpoints were sealed before that cohort's FTS-Diffusion test evaluation. The exploratory-index extension was completed later and did not inform the original Scale Factor selection. Two additional simulated-process fits are described in Appendix I.2.

B.9 PRESPECIFICATION AND RUN ACCOUNTING

Before any evaluation-index test return was scored we fixed the market set, the five training seeds, the three intervals, the Scale Factor equation, the 120-day primary horizon, the 32-component metric taxonomy, and the paired-interval analysis. The primary test of the Scale Factor required all four of the following to hold.

1. The upper bound of the paired 95% interval for the scaled-minus-unscaled discrepancy is below zero at 120 days.
2. The same bound is below zero after equal averaging over 120, 252, and 512 days.
3. Mean tail and volatility discrepancies both decrease.
4. The absolute mean change in the dependence family is at most 5×10^{-12}.

Baseline superiority was deliberately not part of this test, because the question was whether a rule motivated by the exploratory group would reduce the same failure on the evaluation group. The comparisons against every baseline were nevertheless reported for every method, index, horizon, and seed in the study.

The neural market and intervention studies account for 285 trained instances: 180 original exploratory/evaluation fits, 15 white-noise interventions, 30 normalization variants, 30 simulated-intervention fits, and 30 FTS-Diffusion market fits. The 15 classical GJR–GARCH fits and 30

bootstrap outputs are separate. Evaluation training seeds are $\{4101, 4202, 4303, 4404, 4505\}$; exploratory FTS-Diffusion seeds are $\{3101, 3202, 3303, 3404, 3505\}$. All selected samples were finite and passed the degeneracy check. The simulated cross-method comparison uses one replication per method, whereas its memory intervention uses five per process and arm. Analysis timing is summarized in the appendix guide.

C Evaluation Details

C.1 Metric families and scores

Primitive distances. All components compare a generated array with a real one through one of two standardized distances. For finite vectors A and B,

$$\mathcal{W}(A, B) = \frac{W_1(A, B)}{\mathrm{sd}(A \,\|\, B)}, \qquad \delta(a, b; A, B) = \frac{|a - b|}{\mathrm{sd}(A \,\|\, B)},$$

where W_1 is the one-dimensional Wasserstein-1 distance, $\|$ denotes concatenation, and sd is the population standard deviation. $\mathcal{W}$ compares two empirical distributions; δ compares two scalar summaries. Dividing by the pooled standard deviation makes every component unit-free. Block lengths and lags are in trading days.

The six families (32 components).

- *Marginal* (2): $\mathcal{W}$ for pooled daily returns and for their absolute values.
- *Tails* (6): δ for the return quantiles at $u \in \{0.01, 0.05, 0.95, 0.99\}$ and for expected shortfall at $u \in \{0.05, 0.95\}$, i.e., the mean return at or beyond the u-quantile.
- *Dependence* (16): at each lag $\ell \in \{1, 5, 20, 60\}$, the absolute difference between generated and real pooled correlations of $(r_t, r_{t+\ell})$, $(|r_t|, |r_{t+\ell}|)$, $(r_t^2, r_{t+\ell}^2)$, and the leverage pair $(r_t, r_{t+\ell}^2)$. Correlations are invariant to positive rescaling, which gives Equation (8).
- *Volatility* (3): $\mathcal{W}$ between pooled block root-mean-square returns $[b^{-1} \sum_{t=1}^{b} r_t^2]^{1/2}$ with $b \in \{5, 20, 60\}$.
- *Aggregation* (3): $\mathcal{W}$ between pooled non-overlapping b-day return sums for the same three values of b.
- *Path risk* (2): $\mathcal{W}$ for the terminal cumulative log return $\sum_{t=1}^{h} r_t$ and for the maximum drawdown of the cumulative log-return path.

Each family score is the arithmetic mean of its components. The four dependence columns of Table 1 average the four lags of one correlation type; the Dependence column averages all sixteen.

Overall discrepancy. The Overall discrepancy groups the families into four groups and averages the groups with equal weight,

$$J_h = \tfrac{1}{4}\big(Q_{\mathrm{MT},h} + Q_{\mathrm{DV},h} + Q_{\mathrm{A},h} + Q_{\mathrm{R},h}\big),$$

where Q_{MT} pools the 8 marginal and tail components, Q_{DV} pools the 19 dependence and volatility components, Q_{A} the 3 aggregation components, and Q_{R} the 2 path-risk components. It is a fixed study-specific summary chosen before the evaluation, not a proper scoring rule.

Market versus simulated data. For the market comparison, the real array consists of test windows resampled from the 2016–2025 history of the same index, and the score is the standardized discrepancy defined above. For the simulated processes of Appendix I.1, three independent samples of the true process are drawn in addition, and each component is further divided by the median distance among those samples, so that a score of one means "as far from the truth as another true sample." The two scales are therefore not comparable across the two studies. The simulated study also uses horizons of 128, 256, and 512 steps rather than the market horizons, since no trading calendar applies, and caps each horizon score at a fixed value $J_{\mathrm{pen}} \approx 31.995$ chosen before the evaluation, so that a single extreme sample cannot dominate the average (Appendix I.2).

Replications and intervals. Neural methods are run in five paired replications that share the temporal split, test windows, and random streams across methods. For each contrast, the five replication-level differences, each averaged with equal weight over the three evaluation indices, give a paired Student-t 95% interval. These intervals describe fitting, generation, and resampling variability given the observed histories, not uncertainty over future markets.

C.2 Reliability and the score cap

Two safeguards keep a single degenerate sample from dominating a score.

Degeneracy check. Every candidate prefix is screened for near-collapse. Let X be one $N \times h$ prefix, let X^c subtract each path's time mean, and let σ_k be the singular values of X^c. We record the mean absolute pairwise correlation $\bar{c}$ among paths whose temporal standard deviation exceeds machine epsilon, the effective rank

$$R_{\text{eff}} = \exp\Big(-\sum_k p_k \log p_k\Big), \qquad p_k = \frac{\sigma_k^2}{\sum_m \sigma_m^2},$$

and the cross-path dispersion ratio

$$V = \frac{h^{-1}\sum_{t=1}^{h} \text{sd}_i(X_{it})}{\text{sd}_{i,t}(X_{it})}.$$

A prefix is flagged when $\bar{c} \geq 0.99$, $V \leq 0.05$, or $R_{\text{eff}} \leq 1.25$. An all-zero centered panel has effective rank one, and if fewer than two paths have nonzero temporal variation we set $\bar{c} = 1$. If any prefix of a fit is flagged, the penalized summary applies the same penalty at all three horizons. No sample reported in this paper was flagged.

Score cap. For an available horizon aggregate we report $J_h^\dagger = \min(J_h, J_{\text{pen}})$, and a missing aggregate, an unsuccessful fit, or a flagged fit receives J_{pen}. The cap

$$J_{\text{pen}} = 31.994568153915978 = 2 \times 15.997284076957989$$

was fixed before the comparisons reported here, as twice the largest horizon aggregate observed in an earlier 57-run benchmark. It is therefore not estimated from the experiments it summarizes. The cap binds only in the simulated study, where it affects TimeGAN and FTS-Diffusion on the GJR–GARCH process; the market scores never approach it.

C.3 Simulated return processes

The two processes of Appendix I.1 are specified below. All outputs are decimal log returns.

Student-t GJR–GARCH. This process has short volatility feedback with an asymmetric response to negative returns. For $Z_t \overset{\text{iid}}{\sim} t_7$, let $\varepsilon_t = Z_t\sqrt{5/7}$ and

$$r_t = e_t = \sqrt{v_t}\,\varepsilon_t,$$
$$v_t = \omega + \alpha e_{t-1}^2 + \gamma \mathbf{1}\{e_{t-1} < 0\} e_{t-1}^2 + \beta v_{t-1},$$

with $(\omega, \alpha, \gamma, \beta) = (2 \times 10^{-6}, 0.05, 0.08, 0.88)$. We initialize $v_0 = \omega/(1 - \alpha - \gamma/2 - \beta)$ and $e_{-1} = 0$, then discard 2,000 steps.

Long-memory stochastic volatility. This process has slowly decaying volatility dependence and no short feedback term. Let X_t be stationary unit-variance fractional Gaussian noise with $H = 0.78$ and covariance $\gamma_X(k) = \frac{1}{2}(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H})$. Independently, let $Z_t \overset{\text{iid}}{\sim} t_7$ and $\varepsilon_t = Z_t\sqrt{5/7}$. Then

$$V_t = 0.012^2 \exp(0.9X_t - 0.9^2/2), \qquad r_t = \sqrt{V_t}\,\varepsilon_t,$$

so that $\mathbb{E}V_t = 0.012^2$. The fractional Gaussian noise is drawn from its stationary finite-dimensional law, so no burn-in is needed.

D Complete Benchmark Results

D.1 Paired temporal-dependence comparisons

Table 8 provides the paired comparisons supporting the temporal-dependence results in Table 1. All eight intervals lie below zero, supporting the improvement over each baseline across paired replications.

Table 8: Paired comparisons of temporal-dependence fidelity at 120 days on the evaluation indices. Each difference is FinHorizon minus the baseline; negative values favor FinHorizon. Intervals are paired 95% Student-t intervals over five replications, each averaging the three evaluation indices.

Category	Model	Difference	95% CI
Parametric	Student-t GJR–GARCH	-0.0173	$[-0.0281, -0.0065]$
Nonparametric	Stationary bootstrap	-0.0215	$[-0.0274, -0.0155]$
	Circular moving-block	-0.0281	$[-0.0316, -0.0247]$
GAN	SigCWGAN	-0.0252	$[-0.0323, -0.0181]$
	QuantGAN	-0.0132	$[-0.0233, -0.0031]$
	TimeGAN	-0.0479	$[-0.0633, -0.0324]$
Diffusion	Diffusion-TS	-0.0196	$[-0.0240, -0.0151]$
	FTS-Diffusion	-0.0947	$[-0.1414, -0.0479]$

D.2 Results across horizons

Table 9 complements the cross-horizon Dependence comparison in Table 2 with all six metric families and Overall at 512 days. Table 10 reports Overall at all three horizons, paired contrasts, and paired differences in the 120-to-512-day change. Table 11 reports the six metric-family means averaged over all three horizons.

Table 9: Financial scenario fidelity at 512 days. Scores average three evaluation markets and five replications; lower is better. Baselines use their standard outputs; FinHorizon includes its Scale Factor. Bold and underline mark the lowest and second-lowest means among all nine methods, using unrounded scores. Model order matches Tables 1 and 2.

	Return distribution		Temporal structure		Scenario outcomes		Overall
Model	Marginal	Tails	Dependence	Volatility	Aggregation	Path risk	↓
Student-t GJR–GARCH	**0.087**	**0.198**	<u>0.049</u>	**0.177**	**0.186**	0.588	**0.253**
Stationary bootstrap	0.232	0.661	0.056	0.494	0.228	0.505	0.353
Circular moving-block	0.234	0.656	0.059	0.506	0.228	0.547	0.364
SigCWGAN	0.184	0.430	0.061	0.431	0.203	**0.488**	0.295
QuantGAN	0.210	0.467	0.051	0.510	0.319	0.843	0.422
TimeGAN	0.583	0.337	0.085	1.267	0.386	0.820	0.469
Diffusion-TS	<u>0.098</u>	<u>0.230</u>	0.054	<u>0.349</u>	0.197	0.875	0.342
FTS-Diffusion	0.900	2.564	0.146	1.439	0.843	1.490	1.208
FinHorizon	0.129	0.306	**0.038**	0.350	<u>0.192</u>	<u>0.489</u>	<u>0.257</u>

Table 10: Overall discrepancy across horizons on the evaluation indices. Baselines use their standard outputs; FINHORIZON includes its Scale Factor. Panel A reports point estimates averaged over indices and five replications, with the unrounded $512-120$ change in the last column. Panel B reports paired FINHORIZON-minus-baseline contrasts (mean above its 95% Student-t interval; negative favors FINHORIZON) and the paired difference between horizon changes. Bold and underline mark the two smallest Panel-A values per column. Horizon changes are sensitive to fitted drift; Appendix H reports GJR–GARCH drift-removal controls.

Panel A: Point estimates

Model	120	252	512	$512-120$
Student-t GJR–GARCH	**0.184**	**0.206**	**0.253**	0.069
Stationary bootstrap	0.337	0.341	0.353	0.016
Circular moving-block	0.350	0.366	0.364	<u>0.013</u>
SigCWGAN	0.270	0.277	0.295	0.025
QuantGAN	0.345	0.386	0.422	0.077
TimeGAN	0.451	0.455	0.469	0.018
Diffusion-TS	<u>0.234</u>	<u>0.243</u>	0.342	0.109
FTS-Diffusion	1.109	1.149	1.208	0.099
FINHORIZON	0.252	0.261	<u>0.257</u>	**0.005**

Panel B: Paired **FINHORIZON***-minus-baseline contrasts*

Model	120	252	512	Difference in change $(512-120)$
Student-t GJR–GARCH	0.068 $[\ 0.039,\ 0.097]$ -0.085	0.055 $[\ 0.025,\ 0.085]$ -0.080	0.004 $[-0.013,\ 0.021]$ -0.096	-0.064 $[-0.084, -0.043]$ -0.011
Stationary bootstrap	$[-0.138, -0.031]$ -0.098	$[-0.112, -0.048]$ -0.105	$[-0.134, -0.057]$ -0.106	$[-0.052,\ 0.030]$ -0.008
Circular moving-block	$[-0.139, -0.058]$	$[-0.138, -0.073]$	$[-0.128, -0.085]$	$[-0.033,\ 0.018]$
SigCWGAN	-0.018 $[-0.040,\ 0.004]$ -0.093	-0.016 $[-0.048,\ 0.016]$ -0.126	-0.037 $[-0.086,\ 0.011]$ -0.164	-0.019 $[-0.064,\ 0.026]$ -0.071
QuantGAN	$[-0.265,\ 0.079]$ -0.199	$[-0.291,\ 0.040]$ -0.194	$[-0.308, -0.021]$ -0.212	$[-0.136, -0.006]$ -0.013
TimeGAN	$[-0.267, -0.131]$ 0.018	$[-0.258, -0.130]$ 0.018	$[-0.252, -0.171]$ -0.085	$[-0.060,\ 0.034]$ -0.103
Diffusion-TS	$[-0.040,\ 0.076]$ -0.858	$[-0.035,\ 0.071]$ -0.889	$[-0.154, -0.016]$ -0.951	$[-0.150, -0.056]$ -0.093
FTS-Diffusion	$[-0.993, -0.722]$	$[-1.023, -0.754]$	$[-1.086, -0.815]$	$[-0.165, -0.021]$

Table 11: Metric-family decomposition of the Overall discrepancy on the evaluation indices, averaged over three indices, three horizons, and five replications. Lower is better; bold and underline mark the two smallest values among the nine methods.

Model	Marginal	Tails	Dependence	Volatility	Aggregation	Path risk
Student-t GJR–GARCH	**0.099**	**0.216**	0.051	**0.193**	**0.176**	<u>0.421</u>
Stationary bootstrap	0.242	0.678	0.057	0.510	0.230	0.446
Circular moving-block	0.246	0.702	0.062	0.513	0.235	0.485
SigCWGAN	0.188	0.451	0.061	0.434	<u>0.203</u>	**0.413**
QuantGAN	0.216	0.481	<u>0.050</u>	0.520	0.316	0.683
TimeGAN	0.593	0.347	0.085	1.280	0.400	0.751
Diffusion-TS	<u>0.112</u>	<u>0.220</u>	0.056	0.383	0.209	0.582
FTS-Diffusion	0.883	2.499	0.138	1.412	0.832	1.356
FINHORIZON	0.135	0.335	**0.037**	<u>0.358</u>	0.207	0.447

E MECHANISM AND TRAJECTORY EVIDENCE

These analyses retain the original checkpoints, 512-day samples, test splits, panel seeds, and metric weights. The appendix guide records their analysis sequence. Intervals use five seed-level market averages; reference panels are not independent replications. These additional analyses use unadjusted intervals unless stated otherwise.

E.1 EQUAL-LENGTH SEGMENTS ALONG THE GENERATED PATH

We extract days 1–120, 197–316, and 393–512 from each of the same 128 generated paths. Each position is compared with exactly the same three panels of real 120-day test windows, using the original 32-component metric. Consequently, changing the position changes only the generated window. The first-position scores reproduce the original 120-day scores. Standard outputs follow the main horizon table; the memory controls share FINHORIZON's frozen market Scale Factor.

Table 12: Temporal fidelity persists throughout the rollout. Mean discrepancy at three fixed positions, with identical window length and reference data. D denotes Dependence; J denotes Overall. All columns are lower-is-better. The first nine rows are the primary comparison; the final two rows are memory controls.

Model	D: start	D: middle	D: end	J: start	J: middle	J: end
Student-t GJR–GARCH	0.0534	0.0478	0.0541	0.1840	0.1839	0.1979
Stationary bootstrap	0.0575	0.0582	0.0591	0.3367	0.3356	0.3526
Circular moving-block	0.0642	0.0615	0.0593	0.3502	0.3446	0.3237
SigCWGAN	0.0612	0.0590	0.0608	0.2698	0.2825	0.2734
QuantGAN	0.0493	0.0513	0.0549	0.3452	0.3766	0.3795
TimeGAN	0.0839	0.0856	0.0851	0.4509	0.4459	0.4502
Diffusion-TS	0.0556	0.0558	0.0543	0.2336	0.2231	0.2677
FTS-Diffusion	0.1307	0.1606	0.1579	1.1094	1.1481	1.1612
FINHORIZON	0.0361	0.0372	0.0382	0.2519	0.2362	0.2322
White-noise input (M0)	0.0532	0.0537	0.0539	0.2462	0.2430	0.2413
Single-timescale	0.0348	0.0373	0.0369	0.2481	0.2342	0.2321

FINHORIZON has the lowest Dependence mean among the nine primary methods at all three positions. Its end-window Overall mean is 0.2322, compared with 0.2677 for Diffusion-TS and 0.2734 for SigCWGAN. The end-window Dependence interval favors FINHORIZON over seven of eight baselines; the QuantGAN interval includes zero. The paired Overall change from start to end favors FINHORIZON over Diffusion-TS by 0.0538 (interval [0.0054, 0.1022]) and over SigCWGAN by 0.0233 (interval [0.0008, 0.0458]). These comparisons hold observation length fixed and locate the quality advantage within the long rollout.

Table 13: End-window paired differences: FINHORIZON minus comparator. Negative values favor FINHORIZON. The intervals are computed from paired seed differences, not from separate model intervals.

Comparator	Dependence	95% CI	Overall	95% CI
Student-t GJR–GARCH	-0.0159	$[-0.0267, -0.0051]$	0.0343	$[0.0191, 0.0494]$
Stationary bootstrap	-0.0209	$[-0.0298, -0.0121]$	-0.1205	$[-0.1538, -0.0871]$
Circular moving-block	-0.0211	$[-0.0284, -0.0138]$	-0.0916	$[-0.1192, -0.0640]$
SigCWGAN	-0.0226	$[-0.0325, -0.0128]$	-0.0413	$[-0.0701, -0.0125]$
QuantGAN	-0.0167	$[-0.0363, 0.0029]$	-0.1473	$[-0.2954, 0.0008]$
TimeGAN	-0.0469	$[-0.0641, -0.0298]$	-0.2180	$[-0.2713, -0.1648]$
Diffusion-TS	-0.0161	$[-0.0180, -0.0141]$	-0.0356	$[-0.0715, 0.0003]$
FTS-Diffusion	-0.1197	$[-0.1816, -0.0578]$	-0.9290	$[-1.0566, -0.8014]$
White-noise input (M0)	-0.0158	$[-0.0229, -0.0086]$	-0.0092	$[-0.0197, 0.0013]$
Single-timescale	0.0013	$[0.0003, 0.0022]$	0.0000	$[-0.0035, 0.0036]$

E.2 WHERE THE MARKET-LEVEL ADVANTAGES ARISE

The dependence family isolates temporal structure because its correlations are unchanged by positive scale calibration. FINHORIZON has the lowest neural mean in each of the three evaluation markets at the primary horizon. It is first among all methods on DAX and Hang Seng and second to GJR–GARCH on FTSE. These are descriptive mean rankings; paired intervals for each market's strongest neural comparator are shown below.

Table 14: Temporal fidelity by market at 120 days. Dependence discrepancy is scale-invariant; lower is better.

Model	FTSE	DAX	Hang Seng
Student-t GJR–GARCH	0.0368	0.0580	0.0653
Stationary bootstrap	0.0541	0.0461	0.0723
Circular moving-block	0.0609	0.0529	0.0786
SigCWGAN	0.0558	0.0472	0.0807
QuantGAN	0.0652	0.0472	0.0354
TimeGAN	0.1012	0.0821	0.0684
Diffusion-TS	0.0573	0.0553	0.0543
FTS-Diffusion	0.1287	0.1397	0.1237
FINHORIZON	0.0497	0.0326	0.0259

Table 15: FINHORIZON minus the neural baseline with the lowest observed Dependence mean in each market. Comparator selection uses observed means; intervals are descriptive and unadjusted.

Market	Comparator	Difference	95% CI
FTSE	SigCWGAN	-0.0062	$[-0.0176, 0.0053]$
DAX	SigCWGAN	-0.0146	$[-0.0296, 0.0005]$
Hang Seng	QuantGAN	-0.0095	$[-0.0190, 0.0000]$

Scale calibration and Overall rankings have a different market profile. At 120 days the Scale Factor changes FINHORIZON's Overall discrepancy by -0.2058 on FTSE, +0.0108 on DAX, and -0.1545 on Hang Seng. Thus the calibration gains arise mainly on FTSE and Hang Seng, while the temporal advantage appears across all three neural comparisons. For the common-scale Overall comparison, leave-one-market means remain best among neural methods when FTSE or DAX is omitted. When Hang Seng is omitted, SigCWGAN leads and FINHORIZON ranks fourth, third, and second at 120, 252, and 512 days, respectively. The broad claim is therefore cross-market neural temporal fidelity; the average Overall lead reflects the full three-market evaluation.

Table 16: Common-scale neural Overall sensitivity. Each cell gives the model with the lowest mean on the two retained markets, followed by the rank of FINHORIZON among six neural methods.

Omitted market	120 days	252 days	512 days
FTSE	**FINHORIZON** / 1	**FINHORIZON** / 1	**FINHORIZON** / 1
DAX	**FINHORIZON** / 1	**FINHORIZON** / 1	**FINHORIZON** / 1
Hang Seng	SigCWGAN / 4	SigCWGAN / 3	SigCWGAN / 2

E.3 PAIRED MODULE CONTRASTS

Table 17 provides the paired differences and confidence intervals behind the direct-score comparisons in Table 4, together with the 120-day tail and normalization contrasts. Persistent Memory improves Dependence and Volatility at both reported horizons. The normalization variants have near-identical Dependence discrepancies; full-covariance normalization guarantees unit variance for every learned timescale allocation, giving that allocation a controlled interpretation. Tables 19 and 20 provide the complete per-horizon memory and normalization results.

Table 17: Paired module contrasts accompanying Table 4. Each difference is complete FINHORIZON minus the ablated variant, so negative favors FINHORIZON. Intervals are paired 95% Student-t intervals over five replications, each averaging three evaluation indices; differences use unrounded scores. All comparisons are at 120 days unless a 512-day horizon is stated. Dependence is exactly invariant to the Scale Factor. The normalization rows compare the full-covariance readout with diagonal and unnormalized alternatives.

Ablation	Statistic	**FINHORIZON**	Variant	Δ [95% CI]
Scale Factor off	Overall discrepancy	0.252	0.368	−0.117 [−0.139, −0.094]
	Dependence	0.036	0.036	0 (exact)
Memory replaced	Dependence (120 d)	0.036	0.053	−0.017 [−0.020, −0.014]
	Dependence (512 d)	0.038	0.055	−0.017 [−0.021, −0.013]
	Volatility (120 d)	0.366	0.468	−0.102 [−0.180, −0.024]
	Volatility (512 d)	0.350	0.465	−0.115 [−0.166, −0.064]
	Tails (120 d)	0.363	0.270	+0.094 [+0.038, +0.149]
Normalization	Dependence (Full vs. Diagonal)	0.0361	0.0366	−0.0005 [−0.0034, +0.0024]
	Dependence (Full vs. Raw)	0.0361	0.0361	0.0000 [−0.0003, +0.0003]

E.4 THE SCALE FACTOR IN DETAIL

Figure 6 shows the Scale-Factor ablation of Section 4.4 as a horizon profile and as a per-index paired effect. Because the ablated variant differs only by the multiplier c_{val}, the two arms share generated paths, checkpoints, and test windows, so the comparison isolates the factor.

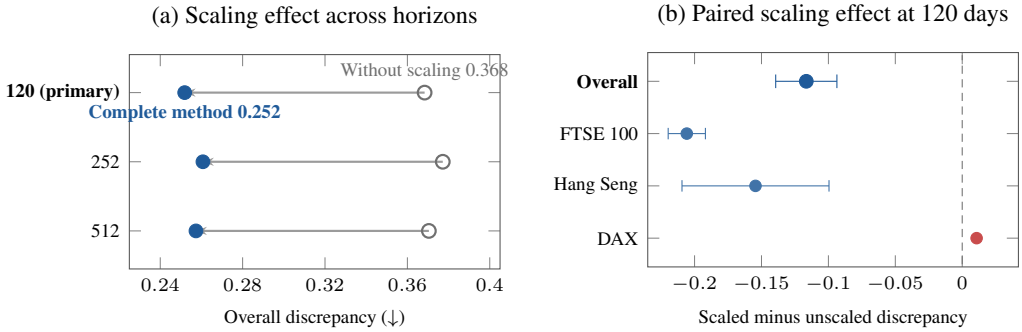


Figure 6: Effect of the Scale Factor on FINHORIZON across the three evaluation indices. The ablated variant sets $c_{\text{val}} = 1$ and the complete model applies c_{val} to exactly the same paths. Panel (a) shows equal-index Overall discrepancy point estimates at the three horizons, with arrows directed from the ablated to the complete output. Panel (b) shows the paired effect at 120 days, computed as scaled minus unscaled discrepancy, with 95% Student-t intervals; negative values favor the factor. No test return enters c_{val}, and positive scaling leaves every dependence statistic unchanged.

E.5 THE MEMORY PATHWAY IN DETAIL

Table 18 decomposes the persistence intervention into all six metric families, averaging the three horizons within each replication. The dependence and block-volatility gains coexist with higher tail discrepancy. Table 17 gives the 120-day contrasts, and Table 19 separates all three horizons.

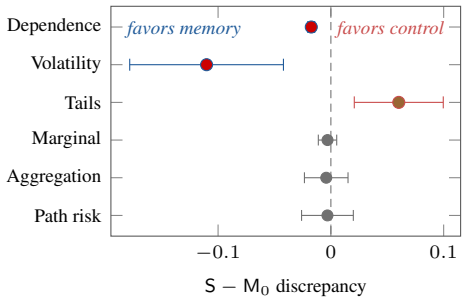
Figure 7 adds the per-family intervals and the timescale allocations that the model actually selects. The allocations are not uniform, which is what makes the normalized mixture interpretable as a choice among known timescales.

Table 19 resolves the same paired intervention by horizon. Dependence and block-volatility gains persist through 512 days. The overall contrasts remain close to zero because improvements in these

Table 18: Persistent Memory (S) against the white-noise input (M_0) by metric family, after equally averaging the three evaluation indices and the three horizons within each replication. Here $\Delta = S - M_0$, so negative favors the memory. The six family contrasts are secondary and unadjusted for multiplicity; lower is better.

Family	M_0	Memory S	Δ	95% CI for Δ
Marginal	0.138	0.135	−0.003	[−0.011, 0.005]
Tails	0.275	0.335	0.060	[0.021, 0.100]
Dependence	0.054	0.037	−0.017	[−0.020, −0.014]
Volatility	0.468	0.358	−0.110	[−0.178, −0.042]
Aggregation	0.211	0.207	−0.004	[−0.023, 0.015]
Path risk	0.450	0.447	−0.003	[−0.026, 0.020]

(a) Selective temporal-family gains



(b) Broad-band allocation without collapse

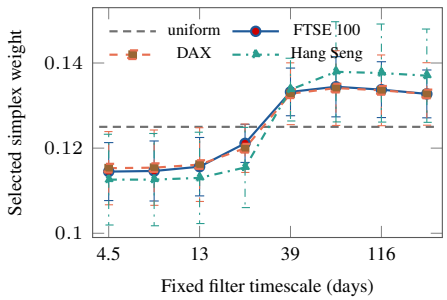


Figure 7: Memory-signal diagnostics on the evaluation indices. Panel (a) reports the paired $S - M_0$ family contrasts averaged over three indices and three horizons; negative values favor the memory, and the six intervals are descriptive and unadjusted for multiplicity. Panel (b) reports index-specific selected simplex allocations over the fixed decay bank, with the dashed line marking uniform allocation. Error bars are 95% Student-t intervals across five training replications.

families coexist with changes elsewhere, including tails. This decomposition identifies the memory's temporal contribution within the complete generator.

Table 19: Persistent Memory improves dependence and block-volatility fidelity across all three horizons. FINHORIZON and the white-noise input control share the market-level Scale Factor and use five paired training replications. Each replication averages the three evaluation indices. Differences are FINHORIZON minus white-noise input; intervals are paired 95% Student-t intervals, unadjusted for the per-family comparisons.

Statistic	Horizon	**FINHORIZON**	White-noise input	Difference	95% CI
Dependence	120	0.0361	0.0532	−0.0172	[−0.0199, −0.0145]
	252	0.0371	0.0550	−0.0179	[−0.0205, −0.0153]
	512	0.0378	0.0548	−0.0170	[−0.0213, −0.0128]
Volatility	120	0.3656	0.4678	−0.1022	[−0.1804, −0.0240]
	252	0.3592	0.4724	−0.1131	[−0.1921, −0.0341]
	512	0.3498	0.4650	−0.1152	[−0.1662, −0.0643]
Tails	120	0.3635	0.2697	+0.0938	[+0.0384, +0.1493]
	252	0.3350	0.2759	+0.0591	[+0.0099, +0.1084]
	512	0.3056	0.2780	+0.0276	[−0.0009, +0.0562]
Overall	120	0.2519	0.2462	+0.0056	[−0.0155, +0.0268]
	252	0.2607	0.2627	−0.0020	[−0.0214, +0.0174]
	512	0.2573	0.2570	+0.0003	[−0.0104, +0.0110]

E.6 MEMORY NORMALIZATION IN DETAIL

The denominator comparison keeps the same correlated states and shared shock in every arm. Full uses $m_t = \pi^\top z_t / \sqrt{\pi^\top \Sigma_z \pi}$, Diagonal replaces the denominator by $\sqrt{\pi^\top \operatorname{diag}(\Sigma_z)\pi}$, and Raw uses $m_t = \pi^\top z_t$. Diagonal therefore ignores cross-covariances only when setting the scale of the scalar mixture, and does not make the state shocks independent.

All arms share data, seeds, training budget, checkpoint rule, and Scale Factor. Raw and Diagonal are function-matched to Full at initialization by rescaling only the direct memory loading and the memory-input column of the recurrent layer, and are then retrained, so the contrasts include differences in optimization and in the range allowed by the bounded direct loading. They are not equal-capacity comparisons. The two co-primary comparisons are Full minus Raw and Full minus Diagonal on 120-day Dependence, adjusted together by Holm's method; Diagonal minus Raw is secondary and the remaining entries of Table 20 are descriptive.

Table 20: Normalization contrasts after applying the same Scale Factor to every arm. Entries are paired differences with 95% Student-t intervals across five seeds, and negative values favor the arm that is named first in the column heading.

Metric ↓	Length	Full − Raw	Full − Diagonal	Diagonal − Raw
Marginal	120	$+.001181\,[-.000499, +.002861]$	$-.002265\,[-.008079, +.003550]$	$+.003446\,[-.003773, +.010665]$
	252	$+.001208\,[-.000534, +.002951]$	$-.002625\,[-.009843, +.004593]$	$+.003833\,[-.004848, +.012515]$
	512	$+.001171\,[-.000505, +.002848]$	$-.001938\,[-.008945, +.005068]$	$+.003110\,[-.005165, +.011385]$
Tails	120	$+.002683\,[-.001987, +.007353]$	$-.003111\,[-.033713, +.027490]$	$+.005795\,[-.025069, +.036658]$
	252	$+.002401\,[-.002597, +.007399]$	$-.002593\,[-.018214, +.013027]$	$+.004995\,[-.011988, +.021977]$
	512	$+.002795\,[-.002940, +.008529]$	$+.000404\,[-.006262, +.007070]$	$+.002391\,[-.004573, +.009354]$
Dependence	120	$+.000001\,[-.000329, +.000331]$	$-.000522\,[-.003418, +.002375]$	$+.000522\,[-.002549, +.003594]$
	252	$-.000080\,[-.000650, +.000489]$	$-.000549\,[-.003512, +.002415]$	$+.000469\,[-.002812, +.003749]$
	512	$-.000038\,[-.000476, +.000399]$	$-.000566\,[-.002649, +.001516]$	$+.000528\,[-.001723, +.002779]$
Volatility	120	$+.002603\,[-.001313, +.006519]$	$-.009514\,[-.030924, +.011896]$	$+.012117\,[-.012560, +.036794]$
	252	$+.002499\,[-.001668, +.006666]$	$-.008079\,[-.032348, +.016191]$	$+.010578\,[-.017217, +.038372]$
	512	$+.002539\,[-.001491, +.006569]$	$-.008180\,[-.032884, +.016523]$	$+.010720\,[-.017249, +.038688]$
Aggregation	120	$+.000648\,[-.000513, +.001810]$	$+.000443\,[-.005407, +.006292]$	$+.000206\,[-.005863, +.006274]$
	252	$+.000688\,[-.000482, +.001859]$	$+.000506\,[-.005271, +.006282]$	$+.000183\,[-.005514, +.005880]$
	512	$+.000809\,[-.000476, +.002094]$	$-.000672\,[-.004435, +.003091]$	$+.001481\,[-.002243, +.005205]$
Path risk	120	$+.000832\,[-.000728, +.002392]$	$-.001411\,[-.008110, +.005289]$	$+.002243\,[-.005207, +.009692]$
	252	$+.001173\,[-.000732, +.003077]$	$-.002982\,[-.008369, +.002406]$	$+.004154\,[-.002688, +.010997]$
	512	$+.000908\,[-.000822, +.002639]$	$+.000194\,[-.004342, +.004731]$	$+.000714\,[-.004073, +.005501]$
Overall discrepancy	120	$+.001050\,[-.000696, +.002796]$	$-.001452\,[-.004984, +.002080]$	$+.002502\,[-.001537, +.006542]$
	252	$+.001073\,[-.000837, +.002983]$	$-.001704\,[-.005177, +.001770]$	$+.002777\,[-.002131, +.007684]$
	512	$+.001119\,[-.000957, +.003195]$	$-.000607\,[-.003244, +.002031]$	$+.001726\,[-.002109, +.005561]$
Overall change	252−120	$+.000023\,[-.000161, +.000206]$	$-.000252\,[-.002876, +.002373]$	$+.000274\,[-.002496, +.003045]$
	512−120	$+.000069\,[-.000305, +.000442]$	$+.000845\,[-.002976, +.004666]$	$-.000777\,[-.004811, +.003258]$

E.7 ACCOUNTING FOR THE MEMORY CONTRIBUTION TO OVERALL

Equation (3) preserves the original four-group weighting. For each family, we multiply the FINHORIZON-minus-control score difference by its exact weight before averaging. The six contributions therefore sum to the observed Overall difference for every seed, market, and horizon. Negative entries lower Overall discrepancy.

Table 21: Original-weight contributions to Overall: FINHORIZON minus white-noise input control. Contributions sum to Overall before rounding.

Contribution	120 days	252 days	512 days
Marginal	0.0001	-0.0002	-0.0005
Tails	0.0176	0.0111	0.0052
Dependence	-0.0036	-0.0038	-0.0036
Volatility	-0.0040	-0.0045	-0.0045
Aggregation	-0.0010	-0.0010	-0.0010
Path risk	-0.0034	-0.0036	0.0048
Overall	0.0056	-0.0020	0.0003

Table 22: Paired intervals for the weighted contributions above. Individual interval endpoints need not sum to the Overall interval because family differences are correlated.

Contribution	120-day 95% CI	252-day 95% CI	512-day 95% CI
Marginal	$[-0.0006, 0.0009]$	$[-0.0009, 0.0005]$	$[-0.0008, -0.0002]$
Tails	$[0.0072, 0.0280]$	$[0.0019, 0.0203]$	$[-0.0002, 0.0105]$
Dependence	$[-0.0042, -0.0030]$	$[-0.0043, -0.0032]$	$[-0.0045, -0.0027]$
Volatility	$[-0.0071, -0.0009]$	$[-0.0076, -0.0013]$	$[-0.0066, -0.0025]$
Aggregation	$[-0.0062, 0.0042]$	$[-0.0064, 0.0044]$	$[-0.0050, 0.0030]$
Path risk	$[-0.0077, 0.0008]$	$[-0.0112, 0.0039]$	$[-0.0018, 0.0114]$
Overall	$[-0.0155, 0.0268]$	$[-0.0214, 0.0174]$	$[-0.0104, 0.0110]$

Memory consistently improves the two temporal families. At 120 days the weighted Tails cost is larger than their combined gain; at 512 days the Tails cost falls but Path risk offsets part of the temporal gain. This accounting explains the near-zero Overall contrast while retaining the persistent Dependence and Volatility improvements that motivate the module.

E.8 Dependence beyond the original lag range

We additionally measure return, absolute-return, and squared-return autocorrelation and leverage at lags 128 and 192 within the 512-day panels. For each statistic, distance is the median absolute candidate–reference correlation difference over the original three 512-day reference panels. The reported mean averages these eight distances, three markets, and five seeds. These extra lags are scored separately and do not enter the original Dependence family or Overall discrepancy.

Table 23: Exploratory dependence at lags 128 and 192. Lower discrepancy is better. The last two columns report FinHorizon minus each comparator; a negative difference favors FinHorizon.

Model	Discrepancy	95% CI	Difference	Paired 95% CI
Student-t GJR–GARCH	0.0194	$[0.0158, 0.0230]$	0.0003	$[-0.0039, 0.0046]$
Stationary bootstrap	0.0192	$[0.0174, 0.0209]$	0.0006	$[-0.0012, 0.0024]$
Circular moving-block	0.0198	$[0.0188, 0.0208]$	0.0000	$[-0.0015, 0.0014]$
SigCWGAN	0.0202	$[0.0181, 0.0223]$	-0.0004	$[-0.0036, 0.0027]$
QuantGAN	0.0199	$[0.0185, 0.0213]$	-0.0002	$[-0.0015, 0.0012]$
TimeGAN	0.0195	$[0.0184, 0.0206]$	0.0003	$[-0.0025, 0.0030]$
Diffusion-TS	0.0238	$[0.0200, 0.0275]$	-0.0040	$[-0.0078, -0.0002]$
FTS-Diffusion	0.0831	$[0.0267, 0.1394]$	-0.0633	$[-0.1205, -0.0061]$
FinHorizon	0.0197	$[0.0177, 0.0218]$	0.0000	$[0.0000, 0.0000]$
White-noise input (M0)	0.0208	$[0.0198, 0.0218]$	-0.0010	$[-0.0028, 0.0007]$
Single-timescale	0.0202	$[0.0192, 0.0212]$	-0.0005	$[-0.0020, 0.0010]$

FinHorizon scores 0.0197, near several classical and neural models. Paired intervals favor it over the two diffusion baselines, while the intervals against the other original baselines and the white-noise input control include zero. This diagnostic supports separating sustained quality along a 512-day trajectory from a claim of unique accuracy at very long lags.

E.9 A learned single-timescale control

The new control replaces the eight-rate readout with one stationary unit-variance AR(1) state, driven by the same Gaussian innovation and supplied to the same recurrent and scale heads:

$$m_{t+1} = \rho m_t + \sqrt{1-\rho^2}\, u_t, \quad \rho = e^{-1/\tau}, \quad m_0 \sim \mathcal{N}(0,1). \tag{9}$$

The trainable log-timescale is bounded by the original grid endpoints, 4.4814 and 199.4996 days, using a sigmoid parameterization. Its initialization is their geometric mean, 29.9005 days. The control receives the same neural initialization, training arrays, 1,000 updates, optimizer settings, training window streams, validation panels, checkpoint selection rule, final sampling seeds, and market

Scale Factors. Fifteen new fits cover the original five seeds and three markets. The implementation retains 3,203 parameter slots for the frozen trainer interface, of which 3,196 are active; seven former mixture-logit slots have zero gradients. Thus this is a matched neural architecture with a smaller active memory parameterization. The selected timescales range from 30.35 to 35.28 days.

The single-timescale control preserves the dependence benefit of a persistent memory signal. Its Dependence discrepancy is 0.0348 at 120 days and 0.0374 at 512 days, compared with 0.0532 and 0.0548 for the white-noise input control. The eight-rate model scores 0.0361 and 0.0378, respectively. Its mean Overall discrepancies are likewise close to the single-timescale control, with paired intervals spanning zero at all three horizons. These observations locate the supported mechanism in the persistence of the memory signal; they do not establish an accuracy advantage for eight decay rates over one.

Table 24: Complete family means for both persistent variants and the white-noise input control. The same market Scale Factors apply to all three. Lower is better; these are direct scores.

Days	Model	Overall	Marginal	Tails	Dependence	Volatility	Aggregation	Path risk
120	**FINHORIZON**	0.2519	0.1388	0.3635	0.0361	0.3656	0.2153	0.3967
	Single-timescale	0.2481	0.1361	0.3527	0.0348	0.3557	0.2135	0.3948
	White-noise input (M0)	0.2462	0.1365	0.2697	0.0532	0.4678	0.2194	0.4105
252	**FINHORIZON**	0.2607	0.1360	0.3350	0.0371	0.3592	0.2136	0.4561
	Single-timescale	0.2585	0.1338	0.3282	0.0367	0.3509	0.2119	0.4561
	White-noise input (M0)	0.2627	0.1390	0.2759	0.0550	0.4724	0.2177	0.4707
512	**FINHORIZON**	0.2573	0.1289	0.3056	0.0378	0.3498	0.1916	0.4893
	Single-timescale	0.2557	0.1271	0.2994	0.0374	0.3439	0.1903	0.4905
	White-noise input (M0)	0.2570	0.1371	0.2780	0.0548	0.4650	0.1957	0.4701

Table 25: Paired differences between memory controls. Eight minus one means FINHORIZON minus single-timescale memory; one minus none means single-timescale memory minus white-noise input. Negative values favor the first named model.

Metric	Days	Eight minus one	95% CI	One minus none	95% CI
Overall	120	0.0038	$[-0.0014, 0.0089]$	0.0018	$[-0.0228, 0.0264]$
	252	0.0022	$[-0.0002, 0.0047]$	-0.0043	$[-0.0225, 0.0139]$
	512	0.0016	$[-0.0019, 0.0051]$	-0.0013	$[-0.0109, 0.0083]$
Marginal	120	0.0027	$[0.0010, 0.0043]$	-0.0004	$[-0.0136, 0.0128]$
	252	0.0021	$[0.0008, 0.0035]$	-0.0052	$[-0.0159, 0.0056]$
	512	0.0018	$[0.0007, 0.0028]$	-0.0100	$[-0.0144, -0.0055]$
Tails	120	0.0108	$[-0.0071, 0.0287]$	0.0830	$[0.0230, 0.1431]$
	252	0.0068	$[-0.0038, 0.0175]$	0.0523	$[0.0085, 0.0960]$
	512	0.0062	$[-0.0046, 0.0170]$	0.0214	$[0.0006, 0.0423]$
Dependence	120	0.0013	$[-0.0003, 0.0028]$	-0.0185	$[-0.0225, -0.0145]$
	252	0.0003	$[-0.0006, 0.0013]$	-0.0182	$[-0.0214, -0.0150]$
	512	0.0004	$[-0.0001, 0.0008]$	-0.0174	$[-0.0219, -0.0130]$
Volatility	120	0.0098	$[0.0025, 0.0171]$	-0.1121	$[-0.1855, -0.0386]$
	252	0.0084	$[-0.0005, 0.0173]$	-0.1215	$[-0.1930, -0.0500]$
	512	0.0058	$[-0.0002, 0.0118]$	-0.1211	$[-0.1671, -0.0751]$
Aggregation	120	0.0019	$[-0.0004, 0.0041]$	-0.0059	$[-0.0279, 0.0160]$
	252	0.0017	$[-0.0010, 0.0045]$	-0.0059	$[-0.0275, 0.0158]$
	512	0.0013	$[-0.0009, 0.0034]$	-0.0054	$[-0.0216, 0.0107]$
Path risk	120	0.0019	$[-0.0071, 0.0110]$	-0.0157	$[-0.0401, 0.0087]$
	252	0.0000	$[-0.0062, 0.0062]$	-0.0146	$[-0.0399, 0.0107]$
	512	-0.0012	$[-0.0099, 0.0075]$	0.0204	$[-0.0010, 0.0418]$

The single-timescale control has lower observed Marginal discrepancy at all horizons and lower Volatility discrepancy at 120 days, with the corresponding eight-minus-one intervals above zero. The

Dependence and Overall intervals include zero at each horizon; this is not an equivalence test. The control is also included in the position and long-lag diagnostics above, so neither its rollout behavior nor its longer-lag results are excluded from the comparison.

E.10 TERMINAL-SEGMENT DEPENDENCE CURVES

Figure 8 decomposes Figure 3 by market. Both use days 393–512 of existing selected paths, all five seeds, and 128 paths per cell. These choices, the four models, and lags 1–60 were fixed before inspecting curves. Correlations pool within-path lagged pairs without smoothing. Real curves are pointwise medians over each market's three original 120-day reference panels.

The main figure averages markets within each seed, then the five seeds; bands show one sample standard deviation across these five curves. The per-market figure omits market averaging. Bands describe replication variability, not uncertainty across markets. The presented curves use the same generated samples. Intermediate lags add interpretation without changing the score; distances between averaged curves need not equal averaged market-level discrepancies.

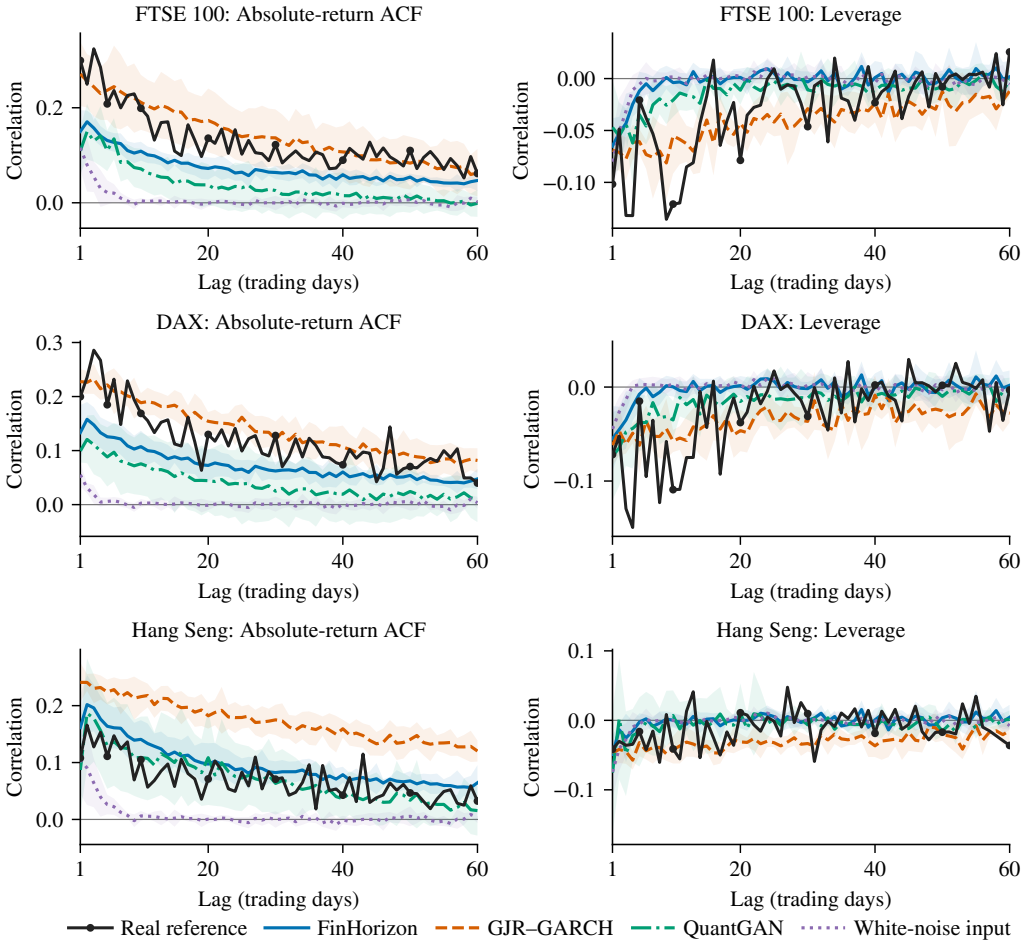


Figure 8: Terminal dependence in all three evaluation markets. Left: absolute-return autocorrelation. Right: return–future-squared-return correlation (leverage). Generated curves show means and one standard deviation across five replications; real curves use the original reference panels. All generated segments are days 393–512.

F ROBUSTNESS TO AGGREGATION, HISTORICAL INFORMATION, AND SV BASELINES

We test sensitivity to metric aggregation, give classical baselines all pre-test observations, and compare closely related SV models. The original splits, samples, and reported results remain fixed; all planned arms are retained. The appendix guide records when each extension was specified, including common-scale SV analysis after unscaled SV results.

F.1 ALTERNATIVE AGGREGATES

We recompute the original candidate scores using equal weights on the six families, equal weights on the 32 components, and the original weights with Path risk removed and the remaining weights renormalized. A separate rank aggregate ranks methods within each seed, market, horizon, and family, then averages the six ranks, markets, and seeds. Lower is better throughout. Rank aggregation gives ordinal importance to each family; it does not make all weighting choices objective. We evaluate both the original outputs and the original common-scale outputs, using the same nine comparison methods. The white-noise input and single-timescale controls remain available in the reproducibility artifact and are not treated as independent external models.

Table 26: FINHORIZON remains competitive under alternative aggregation, while the smallest-increase ranking depends on the aggregate and output protocol. Ranks are among the same nine methods; the last two columns give its 512-minus-120 change and the rank of that change. The rank-aggregate change is in rank units, not discrepancy units.

Aggregate	Rank 120	Rank 252	Rank 512	Change	Change rank
Original outputs					
Original weights	3	3	2	+0.0055	1
Equal family weights	3	3	2	-0.0022	1
Equal component weights	3	3	2	-0.0085	2
Without Path risk	3	3	3	-0.0236	2
Mean family rank	2	2	2	-0.4556	1
Common-scale outputs					
Original weights	4	4	3	+0.0055	2
Equal family weights	4	4	3	-0.0022	2
Equal component weights	3	3	1	-0.0085	2
Without Path risk	4	4	3	-0.0236	2
Mean family rank	3	3	2	-0.2778	4

With original outputs, FINHORIZON ranks second at all three horizons under mean family rank. Removing Path risk changes its 512-day rank from second to third, behind GJR–GARCH and Diffusion-TS, and changes its horizon-increase rank from first to second. The aggregate and its horizon change therefore describe fidelity under the selected weighting, complementing the separate dependence analysis.

We also draw 10,000 six-family weight vectors from a uniform Dirichlet distribution with fixed random seed 20260925. For original outputs, FINHORIZON has the lowest neural mean in 32.94%, 16.69%, and 81.61% of these weight settings at 120, 252, and 512 days; under common scaling the fractions are 98.70%, 98.35%, and 100%. Against all nine methods, the corresponding fractions are only 0.01%, 0.01%, and 13.47% for original outputs, and 0.12%, 1.69%, and 7.92% for common-scale outputs. These fractions describe sensitivity to a chosen distribution over weights, not probabilities that a method is superior. All weights, family scores, and rank conventions are included in the reproducibility artifact.

F.2 GIVING CLASSICAL BASELINES ALL PRE-TEST OBSERVATIONS

We concatenate the original training and validation intervals, 2007–2015, and refit GJR–GARCH or rebuild each bootstrap reference from that complete history. GJR–GARCH retains the original three-start likelihood fitter, five outer seeds, iteration limit, and burn-in; both bootstraps retain 64-day blocks and their original sampling seeds. Outputs receive no further scale multiplier. FINHORIZON retains its original training, selection, and calibration. This comparison gives each procedure access

to the same dated observations, while their fitting objectives and roles for validation differ. As a reproducibility check, train-only reruns exactly recover all 45 original classical candidate arrays before the new-history comparison is evaluated.

Table 27: Access to all pre-test history strengthens the GJR–GARCH Overall baseline. FINHORIZON retains lower Dependence discrepancy at each horizon. Classical procedures fit or resample 2007–2015; FINHORIZON uses its original train/validation fitting and calibration. All scores are lower-is-better.

Model	J_{120}	J_{252}	J_{512}	D_{120}	D_{252}	D_{512}
Student-t GJR–GARCH	0.1618	0.1826	0.2317	0.0514	0.0542	0.0558
Stationary bootstrap	0.2746	0.2728	0.2825	0.0564	0.0544	0.0549
Circular moving-block	0.2882	0.2838	0.2880	0.0599	0.0564	0.0555
FINHORIZON	0.2519	0.2607	0.2573	0.0361	0.0371	0.0378

FINHORIZON's Dependence advantage persists when all three classical procedures use the complete pre-test history: its discrepancy is lower at all three horizons, with all nine paired contrasts supported after Holm correction. GJR–GARCH improves its Overall discrepancies to 0.1618, 0.1826, and 0.2317 and leads FINHORIZON at all three horizons. Table 31 gives the 512-day contrasts.

F.3 ONE- AND TWO-FACTOR STOCHASTIC-VOLATILITY CONTROLS

The likelihood-fitted SV models have one or two independent Gaussian AR(1) log-volatility factors:

$$z_{k,t+1} = \rho_k z_{k,t} + \sqrt{1-\rho_k^2}\, u_{k,t}, \qquad Y_t = \mu_r + \exp\left[\frac{1}{2}\left(\mu_v + \sum_{k=1}^{K} a_k z_{k,t}\right)\right]\epsilon_t. \tag{10}$$

Here $K \in \{1,2\}$, $z_{k,0} \sim \mathcal{N}(0,1)$, and $\rho_k = \exp(-1/\tau_k)$. We fit both Gaussian and variance-one Student-t_5 return innovations. These conventional SV controls have no leverage correlation between return and volatility shocks. The unknowns are the return mean, log-variance mean, and each factor's timescale and stationary log-variance amplitude: four parameters for one factor and six for two. We fit each specification separately on 2007–2013 and 2007–2015, giving 24 fitted market models. Five fixed generation seeds produce 128 continuous-state paths of length 512 per fit; all shorter horizons are prefixes.

Fitting maximizes an approximate marginal likelihood by finite-state integration of the latent AR(1) processes. We use Rouwenhorst transitions, which preserve their conditional moments and stationary variance (Kopecky & Suen, 2010), and a normalized hidden-state forward recursion; the two-factor transition is evaluated as two separable matrix products. Returns are divided by the fitting-history standard deviation for numerical optimization and restored to decimal units for generation. Bounds in these normalized units are $\mu_r \in [-0.5, 0.5]$, $\mu_v \in [-6, 3]$, $\tau_k \in [1, 2000]$ days, and $a_k \in [0.02, 3]$. Each fit starts from three fixed timescale configurations: 5, 30, and 150 days for one factor, and $(3, 30)$, $(10, 100)$, and $(30, 500)$ for two. Initial amplitudes are 0.7 or $(0.5, 0.5)$; the initial log-variance mean matches the normalized empirical variance after the lognormal correction. L-BFGS-B uses at most 300 iterations per start and relative objective tolerance 10^{-9}.

We optimize all starts with 21 states per factor, refine the two best with 41 states, and refine the best with 61. We check the fitted likelihood at 81 states and require a change below 10^{-3} per observation; predefined further refinement to 81 and 101 states is available if needed. An independent continuous-state bootstrap particle filter, with 16,384 particles and four fixed seeds, checks the refined likelihood. Agreement requires the absolute difference to be at most the larger of two log-likelihood units and three estimated Monte Carlo standard errors. All 24 fits pass optimizer, grid, and particle checks. This checks numerical approximation, not global optimality or identification of each individual factor. Final samples use the continuous Gaussian AR(1) model, never the discrete integration grid.

All four fitted-mean SV procedures receive exactly FINHORIZON's market-level multiplier. In this comparison, FINHORIZON scores 0.2573 against 0.3223–0.3325 at 512 days, a reduction of 20.2–22.6%, with all four paired contrasts supported after Holm correction. These are fitted-mean, unlevered SV controls. Appendix G examines how mean specification and volatility dynamics contribute to these scores. Removing the fitted drift substantially reduces the scores, while refitting

Table 28: Training-only SV outputs. Means use the same three markets, five generation seeds, and original scores. J is Overall and D is Dependence; all columns are lower-is-better. FINHORIZON is the unchanged calibrated model.

Model	J_{120}	J_{252}	J_{512}	D_{120}	D_{252}	D_{512}
One-factor Gaussian SV	0.3324	0.3416	0.3905	0.0568	0.0491	0.0495
Two-factor Gaussian SV	0.3119	0.3430	0.3890	0.0529	0.0564	0.0538
One-factor Student-t_5 SV	0.3665	0.3772	0.4182	0.0415	0.0389	0.0386
Two-factor Student-t_5 SV	0.3732	0.3881	0.4208	0.0387	0.0389	0.0390
FINHORIZON	0.2519	0.2607	0.2573	0.0361	0.0371	0.0378

Table 29: SV with the same validation Scale Factor. Means use the same three markets, five generation seeds, and original scores. J is Overall and D is Dependence; all columns are lower-is-better. FINHORIZON is the unchanged calibrated model.

Model	J_{120}	J_{252}	J_{512}	D_{120}	D_{252}	D_{512}
One-factor Gaussian SV	0.2475	0.2645	0.3325	0.0568	0.0491	0.0495
Two-factor Gaussian SV	0.2355	0.2703	0.3271	0.0529	0.0564	0.0538
One-factor Student-t_5 SV	0.2557	0.2701	0.3223	0.0415	0.0389	0.0386
Two-factor Student-t_5 SV	0.2632	0.2801	0.3270	0.0387	0.0389	0.0390
FINHORIZON	0.2519	0.2607	0.2573	0.0361	0.0371	0.0378

Table 30: SV fitted on all pre-test observations. Means use the same three markets, five generation seeds, and original scores. J is Overall and D is Dependence; all columns are lower-is-better. FINHORIZON is the unchanged calibrated model.

Model	J_{120}	J_{252}	J_{512}	D_{120}	D_{252}	D_{512}
One-factor Gaussian SV	0.2922	0.3005	0.3479	0.0529	0.0458	0.0461
Two-factor Gaussian SV	0.2795	0.3069	0.3522	0.0510	0.0532	0.0518
One-factor Student-t_5 SV	0.3359	0.3434	0.3816	0.0402	0.0381	0.0378
Two-factor Student-t_5 SV	0.3431	0.3550	0.3861	0.0379	0.0387	0.0387
FINHORIZON	0.2519	0.2607	0.2573	0.0361	0.0371	0.0378

SV with zero mean yields 0.2448–0.2645. The comparison thus reflects differences in both drift specification and volatility dynamics. The zero-mean and leverage controls help distinguish these effects.

F.4 PAIRED INFERENCE FOR THE COMPARISONS

Each contrast first averages the same three markets within each seed, then forms a Student-t interval from the five seed-level differences. For SV, each market/history/specification has one deterministic fitted model; the five seeds vary generated paths, not fitted histories. The intervals also include FINHORIZON replication variability and are conditional on these observed markets. We apply Holm correction separately within the three classical comparators or four SV comparators, for each protocol, horizon, and metric. Table 31 shows the 512-day Overall and Dependence contrasts; unrounded comparisons for every horizon and family are supplied with the reproducibility artifact.

Table 31: FINHORIZON-minus-comparator differences at 512 days. Negative favors FINHORIZON. Intervals are unadjusted paired 95% intervals; the adjacent p values receive the within-family Holm correction described above.

Comparator	ΔJ	95% CI	p_{Holm}	ΔD	95% CI	p_{Holm}
Classical, train + validation						
Student-t GJR–GARCH	$+0.0257$	$[+0.0118, +0.0395]$	0.020	-0.0180	$[-0.0323, -0.0037]$	0.025
Stationary bootstrap	-0.0251	$[-0.0469, -0.0034]$	0.033	-0.0172	$[-0.0191, -0.0153]$	< 0.001
Circular moving-block	-0.0306	$[-0.0506, -0.0107]$	0.026	-0.0177	$[-0.0224, -0.0130]$	< 0.001
SV, shared Scale Factor						
One-factor Gaussian SV	-0.0752	$[-0.1154, -0.0350]$	0.013	-0.0118	$[-0.0176, -0.0060]$	0.020
Two-factor Gaussian SV	-0.0697	$[-0.1083, -0.0312]$	0.013	-0.0161	$[-0.0243, -0.0078]$	0.020
One-factor Student-t_5 SV	-0.0650	$[-0.0925, -0.0375]$	0.008	-0.0008	$[-0.0044, +0.0028]$	0.974
Two-factor Student-t_5 SV	-0.0696	$[-0.0874, -0.0519]$	0.002	-0.0012	$[-0.0056, +0.0032]$	0.974
SV, train + validation						
One-factor Gaussian SV	-0.0906	$[-0.1323, -0.0489]$	0.005	-0.0084	$[-0.0138, -0.0029]$	0.039
Two-factor Gaussian SV	-0.0949	$[-0.1343, -0.0554]$	0.005	-0.0141	$[-0.0204, -0.0077]$	0.014
One-factor Student-t_5 SV	-0.1243	$[-0.1593, -0.0893]$	0.002	-0.0001	$[-0.0040, +0.0038]$	1.000
Two-factor Student-t_5 SV	-0.1288	$[-0.1479, -0.1096]$	< 0.001	-0.0010	$[-0.0052, +0.0033]$	1.000

G SEPARATING PERSISTENCE, CRITIC INPUTS, AND DRIFT

These controls test the critic's cumulative-return channel, drift, training objective, and leverage while retaining all planned arms. Their sequential specification is recorded in the appendix guide.

All comparisons reuse FTSE, DAX, and Hang Seng, training in 2007–2013, validation in 2014–2015, and testing in 2016–2025. Seeds 4101, 4202, 4303, 4404, and 4505 each supply 128 generated paths of length 512; 120- and 252-day scores use prefixes. Every reported model receives the original market-level Scale Factor. The original score and all 32 components remain fixed. Raw-output scores, market-specific results, window-position scores, and selected and terminal checkpoint scores are included in the reproducibility artifact.

G.1 INTERVENING ON THE CRITIC'S CUMULATIVE-RETURN CHANNEL

The original causal temporal-convolution critic receives four channels: returns, absolute returns, log one plus squared returns, and cumulative returns. We set the fourth channel to zero after the original training-only standardization. Across 15 paired runs, we use the same initialization and training data, changing only this input channel. Channel dimensions, other features, the training budget of 1,000 generator updates, and validation selection rules remain unchanged. The critic still receives complete return sequences, so this intervention tests the contribution of the explicit cumulative-return representation, rather than removing all path information.

Table 32: Critic-channel intervention. All scores are lower-is-better; J denotes Overall, R Path risk, and D Dependence.

Model	J_{120}	J_{252}	J_{512}	R_{512}	D_{512}
FINHORIZON	0.2519	0.2607	0.2573	0.4893	0.0378
Cumulative channel off	0.2465	0.2579	0.2542	0.4919	0.0368

The off-minus-original Overall difference is -0.00314, with a paired 95% interval of [-0.00902, 0.00274]. Path risk improves slightly with the channel present: the off-minus-original difference is 0.00260, with interval [0.00027, 0.00493]. Its Holm-adjusted p value is 0.0723 across the two primary endpoints. The channel's observed weighted Path-risk contribution to Overall is 0.00065, compared with the earlier fitted-mean SV gaps of approximately 0.065–0.075. The 120-to-512-day interaction is also unresolved: 0.00223 for Overall and 0.00706 for Path risk, with both paired intervals spanning zero.

G.2 MATCHING THE RETURN-MEAN SPECIFICATION

FINHORIZON's return equation has zero conditional mean, whereas the original SV fits estimated a daily return mean. First, we subtract that fitted mean from each already generated SV path before applying the shared scale. This intervention leaves all volatility states and random innovations unchanged; it does not center each path by its realized sample mean. It directly tests the effect of extrapolating the fitted drift.

Table 33: Removing fitted drift on the same original SV samples. Entries are 512-day Overall discrepancies; lower is better. Each reduction passes Holm correction across the four specifications.

Model	Fitted drift	Drift removed	Reduction
One-factor Gaussian	0.3325	0.2473	0.0853
Two-factor Gaussian	0.3271	0.2494	0.0777
One-factor Student-t_5	0.3223	0.2728	0.0495
Two-factor Student-t_5	0.3270	0.2668	0.0602

The mean increase from 120 to 512 days falls from 0.064–0.092 to 0.009–0.018 across the four specifications. Thus a substantial part of the earlier SV horizon deterioration can change without changing its volatility dynamics. This finding concerns these SV comparisons, not the unmeasured mechanisms of the other neural baselines.

Second, we refit all four SV specifications with the return mean fixed to zero from the outset. The remaining three or five parameters use the original bounds, multiple starts, Rouwenhorst refinement, and independent continuous-state particle-likelihood checks in Appendix F.3. All 12 market fits pass these checks. Table 35 gives the complete horizon results. Their 512-day Overall contrasts with FINHORIZON all have paired intervals spanning zero (Table 37).

G.3 HOLDING THE SV GENERATOR FIXED ACROSS TRAINING OBJECTIVES

We use the same one-factor Student-t_5 SV equations and parameter bounds for likelihood and adversarial fits. Both adversarial arms start from the same likelihood solution; one retains the cumulative-return channel, and the other masks it. Training uses the original critic backbone, 128-day windows, batch size 24, five critic updates per generator update, WGAN–GP coefficient 10, Adam learning rate 2×10^{-4} with betas $(0.5, 0.9)$, generator gradient clipping at 10, and 1,000 generator updates. Validation at steps 250, 500, and 1,000 selects the earliest checkpoint within 1% of the best validation score. The same random draws are used to evaluate all candidate checkpoints. Final MLE and GAN samples are paired using the same random inputs. The fitted MLE parameters remain unchanged, but its finite-sample scores may differ from those reported earlier because the generated samples differ.

The first comparison permits a learned mean; the second fixes the mean at zero throughout likelihood and adversarial fitting, leaving three trainable parameters. Across all 60 adversarial fits, paired runs use the same initialization, training data, and selection rules, and training diagnostics remain finite. In the zero-mean comparison, the mean is excluded from optimization.

Table 34: Same SV generator, different fitting objectives. All scores are lower-is-better. The final column evaluates the fixed 1,000-update checkpoint; the other columns use the validation-selected checkpoint. MLE has no adversarial checkpoint and its score is repeated.

Mean	Fitting	J_{120}	J_{252}	J_{512}	R_{512}	D_{512}	Final J_{512}
Free	MLE	0.2623	0.2725	0.3277	0.6888	0.0398	0.3277
	GAN, cumulative	0.2530	0.2577	0.2653	0.4980	0.0407	0.2729
	GAN, no cumulative	0.2543	0.2555	0.2616	0.4677	0.0404	0.2655
Zero	MLE	0.2645	0.2730	0.2669	0.4881	0.0402	0.2669
	GAN, cumulative	0.2591	0.2676	0.2615	0.4841	0.0403	0.2615
	GAN, no cumulative	0.2591	0.2679	0.2621	0.4840	0.0409	0.2636

With a free mean, both adversarial arms substantially improve Overall, and both reduce the fitted positive daily drift. With zero mean, the MLE-minus-cumulative-GAN difference is 0.00532, approximately 2%, with unadjusted interval [0.00144, 0.00920] and Holm $p = 0.0570$. The no-cumulative-minus-cumulative GAN difference is 0.00061, with interval [-0.00475, 0.00598]. These controlled comparisons distinguish large drift-sensitive gains from the smaller remaining effect of adversarial refinement under the fixed training budget.

G.4 SV WITH LEVERAGE

We add leverage to all four SV specifications. Let e_t and $v_{k,t}$ be independent standard Gaussian innovations. The return innovation is e_t for Gaussian SV or $e_t\sqrt{(\nu-2)/W_t}$ for Student-t_5 SV, with $W_t \sim \chi^2_\nu$ and $\nu = 5$. The next innovation of the factor with the shorter timescale is $\ell e_t + \sqrt{1-\ell^2}\, v_{k,t}$; the optional slower factor has innovation $v_{k,t}$. This preserves stationary Gaussian AR(1) factors and introduces a signed return–future-volatility link. The mean is freely fitted, and $\ell \in [-0.95, 0.95]$. The coefficient is a correlation of latent Gaussian shocks, not the raw correlation of a Student-t return with the state innovation.

Fitting uses continuous-state particle marginal likelihood with observation-dependent state transitions. For a standardized Student-t residual ϵ, the conditional mixture variable has distribution $W \mid \epsilon \sim \mathrm{Gamma}((\nu+1)/2, 2/(1+\epsilon^2/(\nu-2)))$, using shape–scale notation. Drawing it reconstructs $e = \epsilon\sqrt{W/(\nu-2)}$ for the correlated state transition. This implements the leverage likelihood rather than reusing the independent-innovation forward recursion.

We initialize from the original unlevered fit with leverage 0, -0.3, and -0.6. Each start minimizes the mean negative log likelihood from two fixed 2,048-particle runs with Powell optimization (at most 60 iterations and 1,800 evaluations). The best is refined with two fresh fixed 8,192-particle runs (at most 50 iterations and 1,800 evaluations). Both stages use relative parameter tolerance 0.002 and objective tolerance 10^{-6}. Independent checks use four seeds at 32,768 particles and four at 65,536. Their means must agree within the larger of two log-likelihood units and three combined Monte Carlo standard errors; fitting-objective optimism must be at most the larger of four units and four standard errors. All 12 fits converge and pass. Their independent likelihoods are also no worse than the nested unlevered fits within two log-likelihood units. Analytic constant-volatility checks, two-observation quadrature checks, and the zero-leverage nested check validate the numerical implementation.

Table 35: Complete matched SV scores under the same market-level Scale Factor. Entries average three markets and five replications. J is Overall and D Dependence; all are lower-is-better. The leverage specifications fit the return mean; the zero-mean specifications fix it.

Model	J_{120}	J_{252}	J_{512}	D_{120}	D_{252}	D_{512}
Mean fixed to zero; no leverage						
One-factor Gaussian	0.2346	0.2408	0.2448	0.0510	0.0483	0.0492
Two-factor Gaussian	0.2272	0.2420	0.2470	0.0485	0.0461	0.0501
One-factor Student-t_5	0.2411	0.2606	0.2638	0.0388	0.0396	0.0407
Two-factor Student-t_5	0.2573	0.2780	0.2645	0.0393	0.0412	0.0409
Mean fitted; leverage fitted						
One-factor Gaussian	0.1793	0.1912	0.2293	0.0453	0.0444	0.0441
Two-factor Gaussian	0.2004	0.2049	0.2295	0.0521	0.0493	0.0508
One-factor Student-t_5	0.2198	0.2413	0.2828	0.0337	0.0329	0.0341
Two-factor Student-t_5	0.2342	0.2480	0.2781	0.0391	0.0391	0.0395
FINHORIZON	0.2519	0.2607	0.2573	0.0361	0.0371	0.0378

G.5 EQUAL-LENGTH SEGMENT COMPARISON WITH THE STRONGEST SV CONTROL

We compare FINHORIZON with one-factor Student-t_5 leverage SV, selected for its lowest mean Dependence among the eight SV controls. Existing 128-path samples per market and seed are sliced into days 1–120, 197–316, and 393–512 and compared against Table 3's identical real reference panels. All three markets and five seeds are retained. Recalculation from verified arrays reproduces the stored Dependence discrepancies; no model is refitted or path regenerated.

Table 36: Equal-length segment comparison with one-factor Student-t_5 leverage SV. Means cover three markets and five seeds; lower is better. Differences are FINHORIZON minus SV. Unadjusted 95% Student-t intervals use five paired seed differences after averaging markets and condition on the observed histories.

Metric ↓	Position	SV	**FINHORIZON**	Difference	95% CI
Dependence	Start	0.0337	0.0361	$+0.0023$	$[+0.0003, +0.0043]$
	Middle	0.0325	0.0372	$+0.0047$	$[+0.0019, +0.0074]$
	End	0.0348	0.0382	$+0.0034$	$[-0.0055, +0.0122]$
Overall	Start	0.2198	0.2519	$+0.0320$	$[-0.0047, +0.0687]$
	Middle	0.2279	0.2362	$+0.0082$	$[-0.0173, +0.0338]$
	End	0.2354	0.2322	-0.0033	$[-0.0276, +0.0211]$

SV retains lower Dependence means at all three positions. At the terminal segment, the Dependence difference is $+0.0034$ with interval $[-0.0055, 0.0122]$, while the Overall difference is -0.0033 with interval $[-0.0276, 0.0211]$. None of the six position-by-metric comparisons passes Holm correction. The SV retains competitive terminal fidelity; Table 4 separately isolates the contribution of persistent memory within FINHORIZON.

G.6 PAIRED UNCERTAINTY AND REPRODUCIBILITY

Each contrast averages the three markets within each seed and uses five paired seed differences for a Student-t interval. These intervals are conditional on the observed market histories. SV is fitted once per market and specification; its five seeds vary generated samples, while FINHORIZON and SV–GAN also vary training. Holm correction groups the two primary channel endpoints (512-day Overall and Path risk); their horizon interactions form a separate pair. For each objective comparison, the three pairwise contrasts form a family separately for each 512-day metric. Each set of four SV comparisons forms a family separately by metric. The displayed 95% intervals are unadjusted; there is no global multiplicity guarantee across all diagnostic analyses.

Table 37: FINHORIZON minus matched SV at 512 days. Negative favors FINHORIZON. Intervals are paired, unadjusted 95% intervals; p values are Holm-adjusted within each four-model group and metric.

Model	ΔJ	95% CI	p	ΔD	95% CI	p
Mean fixed to zero; no leverage						
One-factor Gaussian	$+0.0125$	$[-0.0129, +0.0379]$	0.833	-0.0114	$[-0.0162, -0.0066]$	0.008
Two-factor Gaussian	$+0.0103$	$[-0.0088, +0.0295]$	0.833	-0.0123	$[-0.0157, -0.0089]$	0.002
One-factor Student-t_5	-0.0064	$[-0.0290, +0.0161]$	0.833	-0.0029	$[-0.0085, +0.0026]$	0.433
Two-factor Student-t_5	-0.0071	$[-0.0284, +0.0142]$	0.833	-0.0032	$[-0.0092, +0.0028]$	0.433
Mean fitted; leverage fitted						
One-factor Gaussian	$+0.0281$	$[+0.0017, +0.0544]$	0.167	-0.0063	$[-0.0133, +0.0007]$	0.134
Two-factor Gaussian	$+0.0278$	$[-0.0029, +0.0585]$	0.181	-0.0131	$[-0.0203, -0.0059]$	0.029
One-factor Student-t_5	-0.0254	$[-0.0595, +0.0086]$	0.181	$+0.0036$	$[+0.0005, +0.0067]$	0.097
Two-factor Student-t_5	-0.0208	$[-0.0430, +0.0014]$	0.181	-0.0018	$[-0.0044, +0.0008]$	0.134

The artifact retains endpoint and horizon-interaction contrasts with sample hashes. The study contains 15 channel interventions, one original-channel rerun, 60 SV adversarial fits, 12 zero-mean fits, and 12 leverage fits: 100 successful runs across 20 evaluation arms. Only complete validated runs enter the analysis; corrected initialization and serialization failures were rerun.

H MEAN AND SCALE CONTROLS FOR LONG-PATH FIDELITY

These controls retain the three markets, five seeds, 128 paths per market and seed, 512-day length, reference panels, and scoring components. The appendix guide records their specification after the earlier SV results. Reporting families comprise eight GJR–GARCH contrasts per metric and horizon, and all twelve new SV contrasts jointly.

H.1 GJR–GARCH: SEPARATE MEAN AND SCALE INTERVENTIONS

For each saved Student-t GJR–GARCH path r, we evaluate r, $r-\widehat{\mu}$, $c_{\mathrm{val}}r$, and $c_{\mathrm{val}}(r-\widehat{\mu})$. The fitted constant mean $\widehat{\mu}$ is subtracted before scaling; no individual path is demeaned. GJR–GARCH variance updates use innovations around the fitted mean, so this intervention holds the latent variance trajectory and random innovations fixed. We repeat it for training-only and training-plus-validation fits. The common factor is the original frozen market factor in both cases; applying it after a full-history fit is a diagnostic intervention, not a second fitted calibration.

Table 38: GJR–GARCH mean–scale interventions on the same saved paths. Lower scores are better. Change denotes Overall at 512 minus 120 days. Dependence is invariant to positive scaling, but can change under drift removal because some correlations involve absolute or squared returns.

	Overall ↓			Change	Dependence ↓
Intervention	120 days	252 days	512 days	120 to 512	512 days
Fit on 2007–2013					
Fitted mean, original scale	0.1840	0.2057	0.2533	0.0693	0.0491
Drift removed, original scale	0.1891	0.1919	0.1997	0.0105	0.0495
Fitted mean, common scale	0.2040	0.2327	0.2881	0.0841	0.0491
Drift removed, common scale	0.2108	0.2239	0.2322	0.0214	0.0495
Fit on 2007–2015					
Fitted mean, original scale	0.1618	0.1826	0.2317	0.0699	0.0558
Drift removed, original scale	0.1752	0.1885	0.1957	0.0205	0.0561
Fitted mean, common scale	0.2181	0.2272	0.2762	0.0581	0.0558
Drift removed, common scale	0.2337	0.2393	0.2454	0.0117	0.0561
FINHORIZON	0.2519	0.2607	0.2573	0.0055	0.0378

With training-only GJR–GARCH, drift removal lowers 512-day Overall from 0.2533 to 0.1997 and the horizon change from 0.0693 to 0.0105. With the full pre-test history, the corresponding values change from 0.2317 to 0.1957 and from 0.0699 to 0.0205. Dependence remains above FINHORIZON's at every horizon in all eight arms. These interventions identify an accumulated-mean contribution to the aggregate horizon comparison while preserving the distinct temporal-fidelity result.

Table 39: Dependence after GJR–GARCH drift removal. Positive scaling leaves these values unchanged. The three prefix horizons and three equal-length 120-day segments assess different aspects of long-path temporal fidelity.

	Prefix Dependence ↓			Position Dependence ↓		
History / mean	120	252	512	Start	Middle	End
Train / fitted mean	0.0534	0.0520	0.0491	0.0534	0.0478	0.0541
Train / drift removed	0.0537	0.0523	0.0495	0.0537	0.0481	0.0546
Train+val / fitted mean	0.0514	0.0542	0.0558	0.0514	0.0509	0.0601
Train+val / drift removed	0.0518	0.0546	0.0561	0.0518	0.0513	0.0605
FINHORIZON	0.0361	0.0371	0.0378	0.0361	0.0372	0.0382

H.2 ZERO-MEAN LEVERAGE SV REFITS

We fit one-factor Gaussian and Student-t_5 leverage SV with the return mean fixed to zero, completing the combination of drift and leverage controls. The continuous-state particle likelihood and simulator are those of Appendix G; the mean is excluded from optimization. Each market uses its original training history. We initialize from the validated zero-mean, no-leverage fit and start leverage at 0, -0.3, and -0.6. Three 2,048-particle Powell fits are followed by an 8,192-particle refinement, using the original budgets and parameter bounds. Fresh 32,768- and 65,536-particle likelihood estimates validate numerical agreement and check optimism in the fitted objective.

All six fits pass the original convergence and particle-validation gates. Return innovations, factor state, path count, length, scoring references, and market-level calibration follow the earlier protocol.

Table 40: Paired 512-day contrasts: FINHORIZON minus each GJR–GARCH arm. Negative favors FINHORIZON. Intervals use five seed-level differences, each averaging the same three markets. Holm adjusts the eight arms within each metric; uncertainty is conditional on the observed histories.

History / intervention	Metric	Difference [95% CI]	Holm p
Train / Fitted mean, original scale	Dependence	-0.0113 $[-0.0184, -0.0043]$	0.081
	Overall	$+0.0040$ $[-0.0127, +0.0207]$	0.572
Train / Drift removed, original scale	Dependence	-0.0118 $[-0.0189, -0.0046]$	0.081
	Overall	$+0.0577$ $[+0.0377, +0.0777]$	0.011
Train / Fitted mean, common scale	Dependence	-0.0113 $[-0.0184, -0.0043]$	0.081
	Overall	-0.0308 $[-0.0585, -0.0030]$	0.185
Train / Drift removed, common scale	Dependence	-0.0118 $[-0.0189, -0.0046]$	0.081
	Overall	$+0.0251$ $[-0.0058, +0.0560]$	0.347
Train+val / Fitted mean, original scale	Dependence	-0.0180 $[-0.0323, -0.0037]$	0.095
	Overall	$+0.0257$ $[+0.0118, +0.0395]$	0.041
Train+val / Drift removed, original scale	Dependence	-0.0184 $[-0.0327, -0.0040]$	0.095
	Overall	$+0.0617$ $[+0.0390, +0.0844]$	0.012
Train+val / Fitted mean, common scale	Dependence	-0.0180 $[-0.0323, -0.0037]$	0.095
	Overall	-0.0188 $[-0.0505, +0.0129]$	0.525
Train+val / Drift removed, common scale	Dependence	-0.0184 $[-0.0327, -0.0040]$	0.095
	Overall	$+0.0120$ $[-0.0151, +0.0390]$	0.572

The implementation also passes analytic constant-variance likelihood, two-observation quadrature, nested Gaussian no-leverage, stationary-state, and fixed-zero objective checks. Each SV fit is reused across five sampling seeds; FINHORIZON replications also include training variation. No test score selects a fit.

Table 41: Zero-mean one-factor leverage SV and FINHORIZON under the same Scale Factor. Lower is better. Prefix scores use the same 512-day paths and average three markets and five replications.

	Dependence ↓			Overall ↓		
Model	120	252	512	120	252	512
Gaussian, zero mean	0.0454	0.0442	0.0461	0.2354	0.2511	0.2594
Student-t_5, zero mean	0.0323	0.0331	0.0343	0.2817	0.3082	0.3159
FINHORIZON	0.0361	0.0371	0.0378	0.2519	0.2607	0.2573

Table 42: Effect of the common Scale Factor on zero-mean leverage SV. Dependence is unchanged by positive scaling; Overall measures the combined amplitude-sensitive and temporal discrepancies.

	Overall, raw			Overall, common scale		
Model	120	252	512	120	252	512
Gaussian, zero mean	0.3399	0.3541	0.3645	0.2354	0.2511	0.2594
Student-t_5, zero mean	0.4127	0.4401	0.4541	0.2817	0.3082	0.3159

H.3 JOINT TEMPORAL AND OVERALL FIDELITY AT 512 DAYS

Table 46 provides the discrepancies plotted in Figure 4. Each SV family includes a fixed-zero and an estimated return mean. The FINHORIZON generator fixes the conditional log-return mean at zero; its memory mixture and nonlinear return dynamics are learned from training data. Dependence discrepancy summarizes temporal correlations, while Overall discrepancy combines all six metric families with the fixed weights in Equation 3.

Among the two Gaussian configurations, the estimated-mean model has the lower Dependence discrepancy (0.0441). FINHORIZON has discrepancy 0.0378, a reduction of 14.3%. Among the two Student-t_5 configurations, the estimated-mean model has the lower Overall discrepancy (0.2828), compared with 0.2573 for FINHORIZON, a reduction of 9.0%. Both reductions use unrounded means. The estimated-mean Gaussian SV has lower Overall discrepancy than FINHORIZON, and

Table 43: Paired FINHORIZON minus zero-mean leverage SV contrasts. Negative favors FINHORIZON. Holm correction covers all twelve contrasts (two models, three horizons, and two metrics). Intervals condition on the three fixed histories and use the five seed-level market averages.

Control	Metric	Days	Difference [95% CI]	Holm p
Gaussian, zero mean	Dependence	120	-0.0093 $[-0.0143, -0.0043]$	0.067
		252	-0.0071 $[-0.0123, -0.0020]$	0.159
		512	-0.0084 $[-0.0127, -0.0040]$	0.064
	Overall	120	$+0.0164$ $[-0.0046, +0.0375]$	0.577
		252	$+0.0096$ $[-0.0242, +0.0434]$	0.946
		512	-0.0021 $[-0.0242, +0.0201]$	0.946
Student-t_5, zero mean	Dependence	120	$+0.0037$ $[-0.0017, +0.0091]$	0.580
		252	$+0.0040$ $[-0.0003, +0.0083]$	0.436
		512	$+0.0035$ $[-0.0013, +0.0082]$	0.580
	Overall	120	-0.0299 $[-0.0827, +0.0230]$	0.580
		252	-0.0474 $[-0.0961, +0.0012]$	0.430
		512	-0.0586 $[-0.0796, -0.0375]$	0.018

Table 44: Equal-length segment comparison with zero-mean leverage SV. The same real 120-day reference windows evaluate the beginning, middle, and end of 512-day generated paths; lower is better.

	Dependence ↓			Overall ↓		
Model	Start	Middle	End	Start	Middle	End
Gaussian, zero mean	0.0454	0.0473	0.0462	0.2354	0.2567	0.2291
Student-t_5, zero mean	0.0323	0.0355	0.0342	0.2817	0.3145	0.2918
FINHORIZON	0.0361	0.0372	0.0382	0.2519	0.2362	0.2322

Table 45: Fitted parameters of the zero-mean leverage controls. The log-variance intercept is in training-standardized units; τ is in trading days, a is the stationary factor amplitude, and ℓ correlates return and volatility Gaussian shocks. NLL is the mean negative log likelihood over four fresh 65,536-particle evaluations. All six fits pass the optimizer, particle-agreement, and optimism gates.

Market	Innovations	μ_v (norm.)	τ	a	ℓ	NLL
FTSE	Gaussian	-0.4829	40.2958	0.9275	-0.7758	2103.8191
	Student-t_5	-0.2601	49.4399	0.9350	-0.8839	2120.3699
DAX	Gaussian	-0.3415	34.8990	0.8583	-0.7793	2146.1997
	Student-t_5	-0.0760	39.6857	0.8861	-0.8558	2154.1655
Hang Seng	Gaussian	-0.5712	90.6458	0.8891	-0.5774	2041.2845
	Student-t_5	-0.3348	109.6123	0.9192	-0.6728	2054.1203

Table 46: Temporal and overall fidelity of 512-day financial paths. All SV models have one factor and leverage; all models use the same market-level Scale Factor. Mean return is either fixed at zero or estimated from training data. Dependence discrepancy measures temporal relationships; Overall discrepancy combines all six metric families. Lower discrepancies indicate closer agreement with real data. Entries average three markets and five replications. Full multi-horizon results and paired contrasts are in Appendices G and H.2.

Model	Mean return	Dependence discrepancy ↓	Overall discrepancy ↓
FINHORIZON	Fixed at zero	0.0378	0.2573
Gaussian leverage SV	Fixed at zero	0.0461	0.2594
	Estimated	0.0441	0.2293
Student-t_5 leverage SV	Fixed at zero	0.0343	0.3159
	Estimated	0.0341	0.2828

both Student-t_5 configurations have lower Dependence discrepancy. Thus none of these four SV configurations improves both mean discrepancies simultaneously.

Tables 37 and 43 report paired contrasts for the estimated-mean and zero-mean SV controls, respectively. Their intervals condition on the three observed market histories and use five seed-level market averages. Full multi-horizon scores are in Tables 35 and 41.

I Additional Process, Calibration, and Exploratory Results

I.1 Robustness checks

Table 47: FinHorizon achieves the lowest mean excess discrepancy across the two simulated processes and leads all neural baselines in each process. Excess discrepancy is the model's score minus the score of an independent sample from the true process; lower is better. Mean weights the two processes equally. Bold and underline mark only the lowest and second-lowest values in the Mean column among all nine methods.

Category	Model	GJR–GARCH ↓	Long-memory SV ↓	Mean ↓
Classical	Student-t GJR–GARCH	0.155	1.971	1.063
	Stationary bootstrap	0.901	1.031	0.966
	Circular moving-block	0.520	1.062	<u>0.791</u>
Neural	SigCWGAN	1.990	4.450	3.220
	QuantGAN	8.718	2.981	5.850
	TimeGAN	29.812	8.502	19.157
	Diffusion-TS	6.547	7.988	7.267
	FTS-Diffusion	19.527	14.574	17.051
Proposed	**FinHorizon**	0.145	1.169	**0.657**

SV denotes stochastic volatility. Each process score averages 128-, 256-, and 512-step horizons under the common per-horizon score cap; the mean is computed before rounding. Each method has one replication. The independent reference is zero after subtraction. Appendix I.2 specifies the score cap, and Table 48 reports scores before reference subtraction.

Table 48 gives the unshifted scores behind Table 47. Tables 49 and 50 report the common-rescaling check, in which the same validation-derived, index-specific factor is applied to all nine methods without retraining, reselection, or resampling. The FTS-Diffusion extension reuses all 15 selected market samples and the existing index-specific factors. Its original standard-output results remain unchanged, and the extension is reported only as an exploratory sensitivity introduced after the primary evaluation was complete.

Table 48: Scores on the two simulated processes, normalized by the sampling variability of the true process, before subtracting the true-sample score shown in Table 47. One replication per method; lower is better. Scores use the per-horizon cap of Appendix C.2.

Model	GJR–GARCH	Long-memory SV	Mean
True-process sample	1.294	1.268	1.281
Student-t GJR–GARCH	1.449	3.239	2.344
Stationary bootstrap	2.195	2.299	2.247
Circular moving-block	1.815	2.330	2.072
SigCWGAN	3.284	5.718	4.501
QuantGAN	10.012	4.249	7.131
TimeGAN	31.106	9.770	20.438
Diffusion-TS	7.841	9.255	8.548
FTS-Diffusion	20.821	15.842	18.332
FinHorizon	1.439	2.436	1.938

Table 49: Common rescaling of all nine methods on the evaluation indices. The first three columns give the 120-day Overall discrepancy by index, the next three average indices at each horizon, and the last six do both, by metric family; bold and underline mark the two lowest per column.

	120-day score			Mean across indices			Mean by metric family					
Model	FTSE	DAX	H. Seng	120	252	512	Marg.	Tails	Dep.	Vol.	Agg.	Risk
Student-t GJR–GARCH	0.192	**0.180**	0.240	**0.204**	0.233	0.288	**0.112**	0.369	0.051	**0.289**	**0.154**	0.419
Stationary bootstrap	**0.180**	0.331	**0.156**	0.222	**0.221**	**0.230**	0.133	0.352	0.057	0.312	0.154	**0.349**
Circular moving-block	0.194	0.349	0.162	0.235	0.240	0.237	0.137	0.368	0.062	0.320	0.158	0.379
SigCWGAN	0.190	0.292	0.317	0.266	0.282	0.309	0.190	0.407	0.061	0.414	0.233	0.441
Diffusion-TS	0.223	0.300	0.318	0.281	0.316	0.450	0.169	0.492	0.056	0.481	0.227	0.634
FTS-Diffusion	0.946	1.145	1.227	1.106	1.146	1.207	0.890	2.540	0.138	1.412	0.824	1.322
QuantGAN	0.255	0.316	0.320	0.297	0.337	0.375	0.156	0.383	0.050	0.408	0.276	0.636
TimeGAN	0.451	0.429	0.339	0.406	0.418	0.432	0.467	0.416	0.085	0.997	0.331	0.685
FINHORIZON	0.235	0.364	0.156	0.252	0.261	0.257	0.135	**0.335**	**0.037**	0.358	0.207	0.447

Table 50: Effect of common rescaling at 120 days. Left: each method's Overall discrepancy at its original scale and after rescaling, with the paired change (negative favors rescaling); for FINHORIZON the two columns are the Scale-Factor-off ablation and the complete model. Right: complete FINHORIZON minus each rescaled baseline (negative favors FINHORIZON). Brackets are 95% Student-t intervals over five replications.

Model	Original	Rescaled	Change	**FINHORIZON** − rescaled
Student-t GJR–GARCH	0.184 [0.170, 0.198]	0.204 [0.178, 0.230]	0.020 [−0.017, 0.057]	0.048 [0.010, 0.086]
Stationary bootstrap	0.337 [0.312, 0.362]	0.222 [0.208, 0.237]	−0.114 [−0.128, −0.101]	0.030 [−0.012, 0.071]
Circular moving-block	0.350 [0.311, 0.389]	0.235 [0.206, 0.264]	−0.115 [−0.126, −0.104]	0.017 [−0.016, 0.049]
SigCWGAN	0.270 [0.254, 0.286]	0.266 [0.246, 0.287]	−0.003 [−0.017, 0.010]	−0.015 [−0.040, 0.011]
Diffusion-TS	0.234 [0.204, 0.263]	0.281 [0.245, 0.316]	0.047 [0.032, 0.062]	−0.029 [−0.094, 0.037]
FTS-Diffusion	1.109 [0.955, 1.264]	1.106 [0.933, 1.279]	−0.003 [−0.041, 0.034]	−0.854 [−1.007, −0.701]
QuantGAN	0.345 [0.198, 0.493]	0.297 [0.235, 0.359]	−0.048 [−0.144, 0.048]	−0.045 [−0.133, 0.042]
TimeGAN	0.451 [0.397, 0.505]	0.406 [0.366, 0.447]	−0.045 [−0.077, −0.012]	−0.154 [−0.195, −0.113]
FINHORIZON	0.368 [0.317, 0.420]	0.252 [0.219, 0.285]	−0.117 [−0.139, −0.094]	–

I.2 ADDITIONAL PROCESS AND PER-INDEX RESULTS

Process evaluation. The added FTS-Diffusion results use the same 5,000 training and 1,000 validation returns as the existing process comparison, and the same implementation as in the market study, with ten training-only pattern-recognition initializations and candidate checkpoints at 100, 200, and 400 generation epochs. The earliest checkpoint within 1% of the best 128-day validation score is selected, which is checkpoint 400 on both processes. The model receives unstandardized returns and its outputs are not rescaled, and both final samples were fixed before any test score was computed. All generated prefixes are finite and pass the same validity checks as the other methods. Scores average the 128-, 256-, and 512-day horizons; the one-replication comparison remains descriptive.

Common score cap. The process comparison retains the original scoring rule, so each horizon score is capped at $J_{\text{pen}} \approx 31.995$ before averaging the three horizons. For FTS-Diffusion on GJR–GARCH, this reduces the 512-day score from 35.085 to 31.995 and the three-horizon mean from 21.851 to 20.821. This cap also applies to the original TimeGAN results; it is not an FTS-specific adjustment, and reaching the cap does not indicate a failed run. No FTS-Diffusion score is capped on the long-memory process, where every score falls well below the cap.

Per-index results. Table 51 gives the common-rescaling scores for all six neural generators at each index and horizon. For comparison, the standard-output FTS-Diffusion discrepancies at 120, 252, and 512 days are 0.971, 1.006, and 1.059 on FTSE; 1.146, 1.195, and 1.271 on DAX; and 1.211, 1.247, and 1.294 on Hang Seng. These unscaled values underlie the cross-index means in Table 10. The index-level view also shows heterogeneity that the cross-index mean hides. Under common rescaling, SigCWGAN has the lowest discrepancy at every horizon on FTSE and at 120 and 512 days on DAX, and Diffusion-TS is lowest on DAX at 252 days, whereas FINHORIZON is lowest at every horizon on

Hang Seng, where it stays near 0.15 while Diffusion-TS rises from 0.318 to 0.588. These per-index comparisons are descriptive rather than additional independent tests.

Table 51: Mean discrepancy for the six neural generators on each evaluation index across five paired training replications and three horizons under common validation-based rescaling. The same validation-derived index-specific factor is applied to every method. Results for the classical baselines appear in Tables 49 and 50. Bold marks the lowest mean among the six neural generators.

Index	Model	$120\downarrow$	$252\downarrow$	$512\downarrow$
FTSE 100	**FINHORIZON**	0.235	0.249	0.246
	Diffusion-TS	0.223	0.258	0.427
	FTS-Diffusion	0.946	0.981	1.036
	SigCWGAN	**0.190**	**0.197**	**0.217**
	QuantGAN	0.255	0.294	0.309
	TimeGAN	0.451	0.460	0.477
DAX	**FINHORIZON**	0.364	0.390	0.385
	Diffusion-TS	0.300	**0.293**	0.334
	FTS-Diffusion	1.145	1.195	1.270
	SigCWGAN	**0.292**	0.306	**0.320**
	QuantGAN	0.316	0.377	0.418
	TimeGAN	0.429	0.439	0.466
Hang Seng	**FINHORIZON**	**0.156**	**0.143**	**0.141**
	Diffusion-TS	0.318	0.398	0.588
	FTS-Diffusion	1.227	1.263	1.315
	SigCWGAN	0.317	0.344	0.390
	QuantGAN	0.320	0.341	0.397
	TimeGAN	0.339	0.353	0.351

I.3 EXPLORATORY GROUP

The exploratory group (SSE Composite, S&P 500, Nikkei 225) uses the same split and five paired replications. Table 52 reports FINHORIZON without the Scale Factor alongside the five neural baselines. The original comparison contained four baselines. FTS-Diffusion was added later under a separately frozen protocol and did not inform the Scale Factor selection. Its mean 120-day discrepancy was 0.460, with competitive dependence but much larger tail and volatility discrepancies (Figure 9); this gap motivated the Scale Factor. Applying the candidate rules retrospectively identified the validation-to-training ratio of median 20-day rolling volatility as the strongest simple option; it reduced the 120-day mean from 0.460 to 0.367 (paired change -0.093, interval $[-0.125, -0.060]$) with dependence unchanged. Because these indices were used to select the rule, none of these numbers enters the evaluation in Section 4.

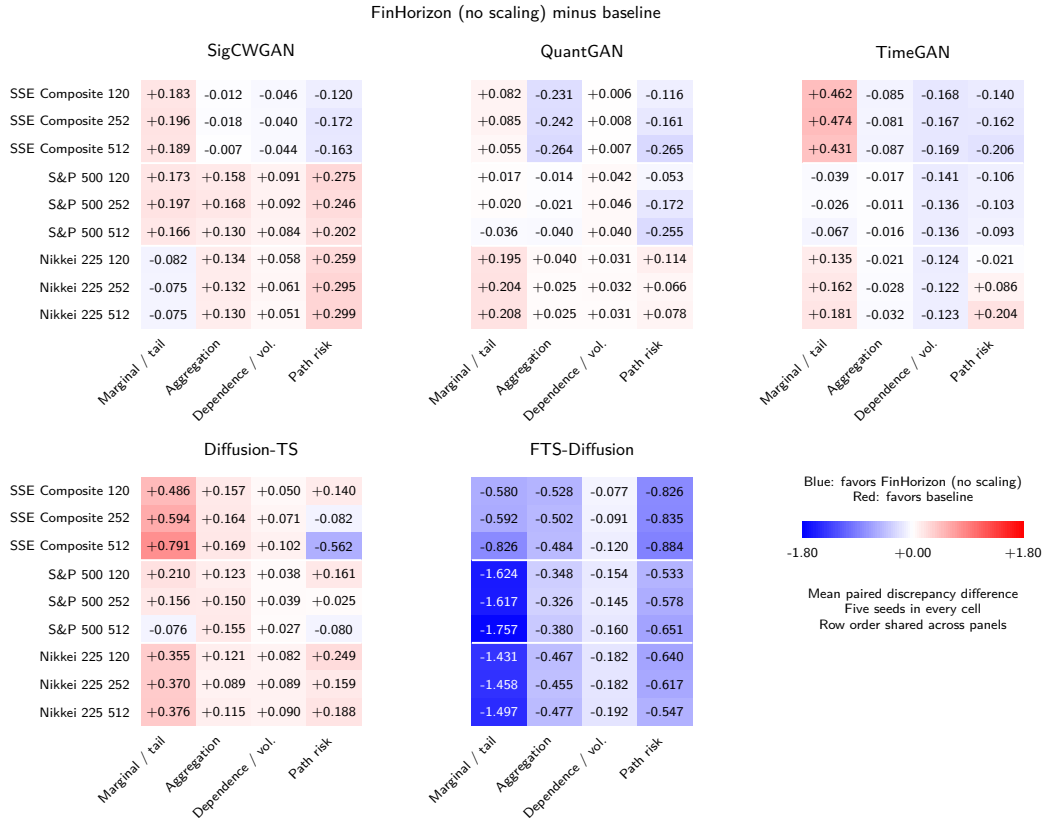


Figure 9: Exploratory indices before the Scale Factor, now including the supplemental FTS-Diffusion comparison. Each cell is the mean paired discrepancy of FINHORIZON without scaling minus a baseline, by metric group, over five replications; blue favors FINHORIZON. The original four-baseline comparison motivated the Scale Factor; the later FTS-Diffusion extension did not inform its selection. All five panels share the same row order and colour scale.

Table 52: Overall discrepancy on the exploratory indices for FINHORIZON without the Scale Factor and for the standard outputs of the neural baselines. The original four-baseline comparison motivated the Scale Factor; the later FTS-Diffusion extension is supplemental. Neither is part of the held-out evaluation.

Index	Model	120	252	512
SSE Composite	**FINHORIZON** (no scale)	0.621	0.641	0.634
	SigCWGAN	0.620	0.650	0.640
	QuantGAN	0.686	0.719	0.751
	TimeGAN	0.604	0.625	0.642
	Diffusion-TS	0.413	0.455	0.509
	FTS-Diffusion	1.124	1.146	1.212
S&P 500	**FINHORIZON** (no scale)	0.370	0.394	0.412
	SigCWGAN	0.195	0.218	0.267
	QuantGAN	0.372	0.426	0.485
	TimeGAN	0.445	0.463	0.490
	Diffusion-TS	0.236	0.301	0.406
	FTS-Diffusion	1.034	1.060	1.149
Nikkei 225	**FINHORIZON** (no scale)	0.389	0.420	0.472
	SigCWGAN	0.296	0.317	0.371
	QuantGAN	0.293	0.338	0.386
	TimeGAN	0.396	0.395	0.414
	Diffusion-TS	0.187	0.243	0.280
	FTS-Diffusion	1.069	1.098	1.151

J Extended Related Work

Neural time-series generators. Most recent work on financial time-series generation trains one neural network to reproduce the whole distribution of return paths. GAN-based generators differ mainly in how they compare generated and real paths: TimeGAN adds a supervised loss on latent recurrent dynamics (Yoon et al., 2019), QuantGAN uses temporal convolutions to capture long-range dependence (Wiese et al., 2020), and SigCWGAN, COT-GAN, and PCF-GAN compare path distributions through signatures, causal optimal transport, and characteristic functions (Liao et al., 2024; Xu et al., 2020; Lou et al., 2023). Diffusion-based generators such as Diffusion-TS (Yuan & Qiao, 2024) and general-purpose sequence generators (Jeha et al., 2022; Naiman et al., 2024; Jarrett et al., 2021; Zhou et al., 2023) follow the same idea with different backbones, and TSGBench standardizes their evaluation (Ang et al., 2023). FinHorizon brings explicit persistent dynamics into this learning framework. Its memory carries latent shocks across multiple timescales, while a separate factor calibrates the scale of complete paths. Our evaluation tests both the learned temporal structure and the quality maintained as generation extends to 512 days.

Multi-timescale structure of volatility. That volatility persists over several horizons at once is well established. GARCH-type models capture short persistence with a one-step recursion (Bollerslev, 1986; Glosten et al., 1993); heterogeneous-horizon models such as HAR show that combining daily, weekly, and monthly components already explains much of the slow decay of volatility dependence (Ding et al., 1993; Baillie et al., 1996; Corsi, 2009); and rough-volatility models with their Markovian approximations show that a handful of exponential components reproduces this decay over a finite range of lags (Gatheral et al., 2018; Abi Jaber & El Euch, 2019b;a), an observation that goes back to the aggregation of short-memory processes (Granger, 1980). These results motivate explicit representations of persistence. FinHorizon uses that principle inside a neural path generator: its memory receives the generator's own shocks, and its timescale weights are learned from return-path distributions. Multi-timescale recurrent networks and structured state-space layers likewise use mixtures of decay rates to carry information over long sequences (Challet & Ragel, 2024; Gu et al., 2022; Orvieto et al., 2023; Miller & Hardt, 2019). Our construction combines a stationary shared-shock memory, a nonlinear return generator, and independent scale calibration. The exact stationary covariance gives the memory readout unit variance for every learned allocation over timescales.

Separating scale from structure. The view that magnitude and temporal structure are different questions appears in several forms. Copula-based time-series models separate the marginal distribution from serial dependence (Chen & Fan, 2006), and block bootstraps preserve local ordering by resampling blocks of real data (Künsch, 1989; Politis & Romano, 1991; 1994). Econometric studies further show that shifts in the overall volatility level can be mistaken for long memory (Granger & Hyung, 2004; Mikosch & Stărică, 2004), which is one reason not to learn the two as one. Among neural generators, FTS-Diffusion models recurring patterns with varying durations and magnitudes (Huang et al., 2024), and CoFinDiff conditions generation on a user-specified trend and volatility (Tanaka et al., 2025). These methods control magnitude at the segment level or through a condition supplied to the trained generator. FinHorizon separates scale at the market level. A single factor, estimated from data before the test period, adjusts a frozen generator and stays fixed across all paths and horizons. Our contribution is the combination of this calibration with explicit persistent memory in a learned path generator. Applying the same factor to all baselines provides an additional comparison in which FinHorizon has the best neural scores at all three horizons.