OUT-OF-DISTRIBUTION-GATED MOMENTUM: CONTROLLING EXPOSURE VIA FEATURE-BASED WASSERSTEIN DISTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Standard trend-following strategies rely on volatility scaling to size positions, inherently ignoring the knowledge about uncertainty when deployed in out-of-distribution (OOD) market conditions. We propose a distribution-aware familiarity score to explicitly quantify this uncertainty without relying on learned representations. Our method measures the optimal-transport distance between current market return distributions and their closest analogues in the training data, acting as a smooth, continuous throttle on position exposure. By isolating this sizing mechanism from the directional signal, we guarantee that performance differences stem purely from OOD exposure control. Across five asset classes and multiple validation schemes, our optimal-transport throttle consistently reduces portfolio drawdowns, with ablations confirming the mechanism captures OOD dynamics distinct from standard volatility. We also report critical negative results: jointly learning direction and size fails to converge, and enforcing strict statistical calibration actively degrades performance-highlighting a fundamental tension in OOD detection. Finally, we demonstrate that distribution-aware familiarity serves as a robust, interpretable principle for controlling exposure under domain shift.

1 INTRODUCTION

A trend-following position on a single instrument requires two decisions: *which way* to trade, and *how much* to risk. Direction has a long, well-studied answer - time-series momentum is among the most robust and heavily documented anomalies in empirical finance (Moskowitz et al., 2012). Magnitude is usually answered by volatility scaling: size the position inversely to recent realized volatility (Lim et al., 2019; Moreira & Muir, 2017). That rule has no notion of whether the market condition it is currently sizing into looks anything like the conditions the strategy's assumptions were built on.

This paper asks a narrow question: can an out-of-distribution (OOD) signal, computed with no learned representation at all, adjust position magnitude in a way that adds value beyond volatility scaling? We build the simplest defensible instantiation, T_1 - a k-nearest-neighbour Wasserstein-2 distance in a small, hand-engineered condition-feature space (Figure 1) - and test it as hard as we can: under walk-forward validation and cluster-robust re-analysis, across T_1's own hyperparameters, an alternative direction signal and a 24-year evaluation window, and head-to-head against eleven alternative gating signals built specifically to stress-test the claim. The result is a mechanism with genuine, independently-verified robustness on its primary claim (drawdown reduction), reported alongside a complete account of where simpler or differently-motivated alternatives are competitive.

Design choices. Direction is a classical, non-learned time-series-momentum signal, held fixed across every magnitude-mechanism comparison (except the learned-direction and alternative-direction ablations), so any change in outcome is attributable to the magnitude mechanism alone. T_1 itself uses no learned encoder: it is a distance computed directly on the trailing return distribution, indexed by kNN search in an engineered feature space, avoiding the failure modes - unstable training, representation collapse - a learned encoder would add on top of an already-hard magnitude-sizing problem. We report our own encounter with that class of failure directly: a Sharpe-ratio-loss

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.

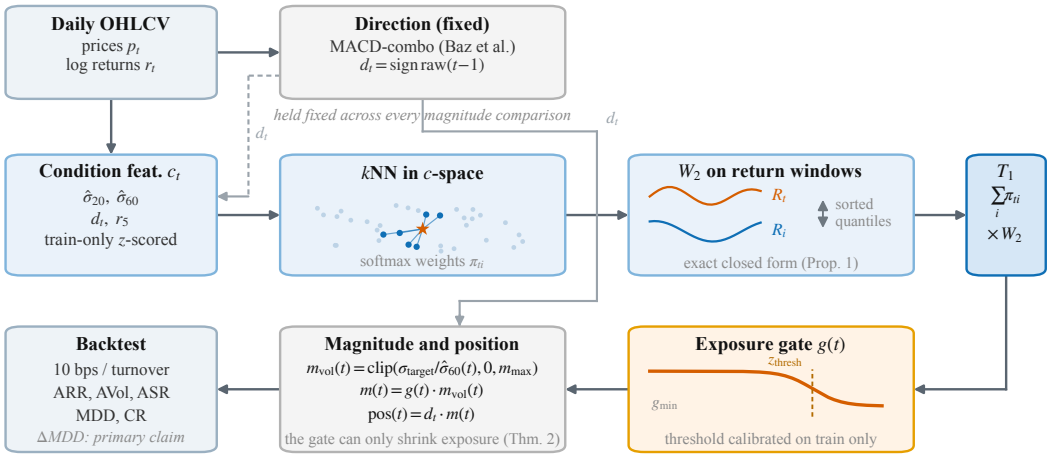


Figure 1: The mechanism. Direction is fixed and non-learned, held constant across every magnitude comparison. The magnitude branch scores each day's familiarity by finding its nearest neighbours in an engineered condition-feature space and measuring the Wasserstein-2 distance between the query's trailing return window and each neighbour's. The score T_1 passes through a train-calibrated sigmoid gate that can only *shrink* the volatility-scaled magnitude (Theorem 1). Every competing gate in Section 4 replaces the score and nothing else.

network attempting to jointly learn direction and magnitude failed to converge to a non-degenerate solution across six training variants (Appendix S).

Causal timing, verified rather than asserted. A backtest of this kind lives or dies on whether every signal at index t is computable from information available before day t opens. We do not assert that convention, we verify it signal by signal with a controlled perturbation test: change one day's price, and confirm that no same-indexed signal moves. Every signal used in this paper - T_1's inputs, the traded direction, and each of the eleven competing gates - passes that test, and we give the protocol in full so it can be re-run (Appendix B).

Contributions.

- **A magnitude-sizing mechanism**, T_1-gated exposure, that reduces maximum drawdown on 13 of 15 instruments across five asset classes (mean -6.42 pp), validated under walk-forward evaluation (Wilcoxon $p < 10^{-11}$, cluster-robust) and stable across T_1's own hyperparameters, an alternative direction signal (15/15 improved), an extended 2000-2024 window (11/12), and a 5-50 bps cost sweep.
- **A decisive answer to the sharpest objection.** Despite T_1 correlating strongly with realized volatility, a naive volatility-ratio gate with identical machinery improves only 7 of 15 instruments - and residualising volatility out of T_1 entirely *improves* the gate to 14/15.
- **A causal-timing verification protocol** applied to every signal in the paper, ours and each competitor's alike: a controlled perturbation test that establishes, rather than assumes, that no signal at index t depends on day t's close (Appendix B).
- **The most complete competing-signal comparison in this line of work**: eleven alternative gates sharing T_1's exact machinery plus a pre-specified feature-space sweep, reported in full including where competitors exceed T_1.
- **Two tensions made explicit rather than hidden**: restoring nominal calibration coverage destroys the drawdown benefit, and decorrelating the gate across a portfolio removes the benefit with the correlation. Both are characterised as continuous frontiers (Section 4.8).
- **Machine-checked mathematics.** Every mathematical claim in the paper is formalised and verified in Lean 4 with Mathlib (de Moura & Ullrich, 2021; The mathlib Community, 2020), with no unproved steps (Appendix A).

2 Related Work and Background

Time-series momentum and deep momentum networks. Time-series momentum is documented across asset classes and decades (Moskowitz et al., 2012); Baz et al. (2015) propose the multi-timescale MACD-combo signal we adopt as our fixed direction signal. Lim et al. (2019) propose Deep Momentum Networks (DMN), which jointly learn direction and volatility-scaled size via an LSTM trained end-to-end on a differentiable Sharpe ratio. DMN's volatility-scaling baseline is the reference point our gate is layered on and compared against throughout; unlike DMN we deliberately do not learn direction, and our attempt to reproduce DMN's joint learning is reported as a negative result (Appendix S).

Out-of-distribution detection and Wasserstein geometry. T_1 is motivated by a Wasserstein-2 OOD decomposition for conditional generative models (Pham, 2026), which separates generalisation error under distribution shift into an interpolation-sparsity term (distance from training support), a geometric-distortion term, and an in-distribution fitting-error term. That work requires a trained encoder to define the embedding space; this paper specialises the interpolation-sparsity term to a setting with *no learned encoder at all*. Background on optimal transport follows Villani (2008) and Santambrogio (2015); the energy distance (Székely & Rizzo, 2013), maximum mean discrepancy (Gretton et al., 2012) and dynamic time warping (Sakoe & Chiba, 1978) that we test as alternative inner distances are all classical.

Change-point detection and turbulence indices. Faber et al. (2021) (WATCH) detect change points via Wasserstein distance between sliding windows; Kritzman & Li (2010) propose the Mahalanobis financial turbulence index; a Markov-switching volatility model (Hamilton, 1989), a GARCH(1,1) model (Engle, 1982; Bollerslev, 1986) and a two-sided CUSUM statistic (Page, 1954) are three further classical regime-aware alternatives. We implement and compare against all of these directly rather than only citing them (Section 4.5). Wood et al. (2022) and Wood et al. (2021) instead insert change-point detection and attention into a Deep Momentum Network; both remain end-to-end learned architectures, distinct from this paper's encoder-free design.

Three further strands are directly relevant and are discussed in full in Appendix E: classical volatility timing (Moreira & Muir, 2017), which our volatility-ratio ablation (Section 4.5) tests T_1 against directly; conformal risk control (Vovk et al., 2005; Gibbs & Candès, 2021), whose spirit our train-calibrated threshold follows without claiming a formal guarantee - and Section 4.8 shows directly why we cannot claim one; and the large regime-aware and mixture-of-experts trading literature (Shi et al., 2025; Vallarino, 2025; Ding et al., 2024; Zhang et al., 2025; Feng et al., 2019; Gao et al., 2023; Li et al., 2024), which routes or predicts *direction* conditioned on regime where our contribution is orthogonal - we hold direction fixed and study only whether an OOD-motivated distance is a useful *magnitude* signal. Appendix F explains, method by method, why none of the latter is a controlled comparison for our task, and why we chose not to import their reported numbers.

3 Method

3.1 Direction: fixed time-series momentum

Every experiment other than the learned-direction and alternative-direction ablations uses the same, non-learned direction signal. Given daily log returns r_t and prices p_t, we compute a multi-timescale MACD-combo signal following Baz et al. (2015): for kernel pairs $(S, L) \in \{(8, 24), (16, 48), (32, 96)\}$,

$$q_{S,L}(t) = \frac{\mathrm{EWMA}_S(p_t) - \mathrm{EWMA}_L(p_t)}{\mathrm{std}_{63}(p_t)}, \qquad \hat{q}_{S,L}(t) = \frac{q_{S,L}(t)}{\mathrm{std}_{252}(q_{S,L})}, \tag{1}$$

combined as $\mathrm{raw}(t) = \tanh\big(\frac{1}{3}\sum_{S,L} \hat{q}_{S,L}(t)\big)$.

Causal timing. $\mathrm{EWMA}_S(p_t)$ incorporates p_t itself - the same closing price that determines the return realised on day t - so using $\mathrm{raw}(t)$ directly would use information not available until day t is over. We therefore define $d_t = \mathrm{sign}(\mathrm{raw}(t-1))$, shifting by one day so direction depends only on prices through the close of day $t-1$, matching the timing $\hat{\sigma}_{20}$, $\hat{\sigma}_{60}$ and m_{vol} already have by construction. Verified by perturbation test (Appendix B). This shifted signal is then held fixed across every condition that uses it.

3.2 MAGNITUDE BASELINE AND THE T_1 SCORE

The baseline magnitude is standard volatility targeting, $m_{\text{vol}}(t) = \text{clip}(\sigma_{\text{target}}/\hat{\sigma}_{60}(t), 0, m_{\max})$ with $\sigma_{\text{target}} = 15\%$ and $m_{\max} = 2$ - the same mechanism used as a baseline in Lim et al. (2019), unmodified, only gated.

Condition features. For each day t we build $c_t = (\hat{\sigma}_{20}(t), \hat{\sigma}_{60}(t), d_t, r_5(t))$ - realized volatility at two horizons, the shifted direction signal, and the trailing 5-day return, all through $t-1$ - z-scored using *only* training-period statistics (train-only standardisation, Appendix C).

Nearest-neighbour weighting. For query day t, let $\mathcal{N}_k(t)$ be its $k = 20$ nearest neighbours in $\{c_i\}$ restricted to the training period, under Euclidean distance d_{ti}, with softmax weights $\pi_{ti} = \exp(-d_{ti}/\tau)/\sum_j \exp(-d_{tj}/\tau)$ and $\tau = 0.5$. The parameters k, τ and the window w below are swept in Section 4.4.

Wasserstein distance. For query window $R_t = \{r_{t-w}, \ldots, r_{t-1}\}$ with $w = 20$ and neighbour window R_i, both scalar samples of equal size, the Wasserstein-2 distance is available in closed form as a root-mean-square difference of order statistics, and we define

$$T_1(t) = \sum_{i \in \mathcal{N}_k(t)} \pi_{ti}\, W_2(R_t, R_i), \qquad W_2(R_t, R_i) = \left(\tfrac{1}{n}\textstyle\sum_j \left(R_t^{(j)} - R_i^{(j)}\right)^2\right)^{1/2}, \tag{2}$$

where $R^{(j)}$ denotes the j-th order statistic. The closed form holds because the optimal coupling on the real line is the monotone rearrangement, which we state and machine-check:

Proposition 1 (1-D Wasserstein-2 closed form; Lean: `sqCost_le_sqCost_comp_perm`)**.** *Let $x, y : \{1, \ldots, n\} \to \mathbb{R}$ be non-decreasing. Then for every permutation σ, $\sum_j (x_j - y_j)^2 \le \sum_j (x_{\sigma(j)} - y_j)^2$. Consequently, for two empirical measures on $\mathbb{R}$ with n equally-weighted atoms each, W_2^2 equals $\frac{1}{n}\sum_j (x_{(j)} - y_{(j)})^2$, with no projection or Monte-Carlo approximation.*

The proof, its Lean formalisation and the complete list of machine-checked statements are in Appendix A; the unequal-length code path is documented in Appendix D.

Remark (informal). T_1 is a plug-in kNN estimator of a population W_2 distance between conditional return distributions; standard kNN consistency results (Biau & Devroye, 2015) suggest concentration as $k, n_{\text{train}} \to \infty$ *under stationarity* - an assumption Section 4.8's coverage check shows does not hold here. This motivates the construction; it is not a claimed theorem.

3.3 THE EXPOSURE GATE

Threshold calibration uses only the training period, self-excluded. Let μ, σ^2 be the training-period mean and variance of T_1 and let z_{thresh} be the z-score of its training-period 90^{th} percentile. The gate is

$$g(t) = g_{\min} + (1 - g_{\min}) \cdot \frac{1}{1 + \exp\big(s\,(z_t - z_{\text{thresh}})\big)}, \qquad z_t = \frac{T_1(t) - \mu}{\sigma}, \tag{3}$$

with defaults $s = 2.5$, $g_{\min} = 0.2$; the final magnitude is $m(t) = g(t)\, m_{\text{vol}}(t)$ and the position is $d_t\, m(t)$. A hard variant, $g_{\text{hard}}(t) = \mathbb{1}[T_1(t) \le \mu + z_{\text{thresh}}\sigma]$, is also evaluated (Appendix Figure 16(a)). *This identical mechanism, applied to a different scalar score in place of $T_1(t)$, is what every competing gate in Section 4 uses - the only thing that ever changes is the score.*

Three properties make the comparisons in Section 4 interpretable; all are machine-checked (Appendix A, Table 2).

Theorem 1 (Gate range, monotonicity, contraction)**.** *For $0 \le g_{\min} \le 1$ and $s > 0$: (i) $g(t) \in [g_{\min}, 1]$; (ii) g is strictly decreasing in $T_1(t)$, so a more out-of-distribution day never receives more exposure; (iii) $g(t)\, m_{\text{vol}} \le m_{\text{vol}}$ for any $m_{\text{vol}} \ge 0$ - the gate can only shrink exposure, never lever it up. Moreover $g = \frac{1+g_{\min}}{2}$ exactly at the calibrated threshold and $g \to g_{\min}$ as $s \to \infty$ above it, recovering the hard variant.*

Corollary 1 (Any Sharpe difference is a *timing* effect)**.** *The annualised Sharpe ratio is invariant under multiplying a return series by a positive constant, so a gate applying a* constant *factor would leave it exactly unchanged. Every Sharpe difference we report is attributable to the gate varying over time, not to it reducing average exposure.*

Table 1: Mean risk-adjusted return and drawdown, 15 instruments, single 60/40 split. Per-instrument breakdown: Appendix G.

Method	Mean ASR	Mean MDD (%)	MDD improved
Volatility scaling (baseline, Lim et al., 2019)	−0.007	31.13	-
T_1-gate, smooth (ours)	−0.029	**24.72**	**13 / 15**
T_1-gate, hard breaker (ablation)	0.005	26.14	10 / 15
Learned direction, logistic (§R.1)	−0.534	37.75	-
Learned direction, MLP (§R.1)	−0.253	33.82	-

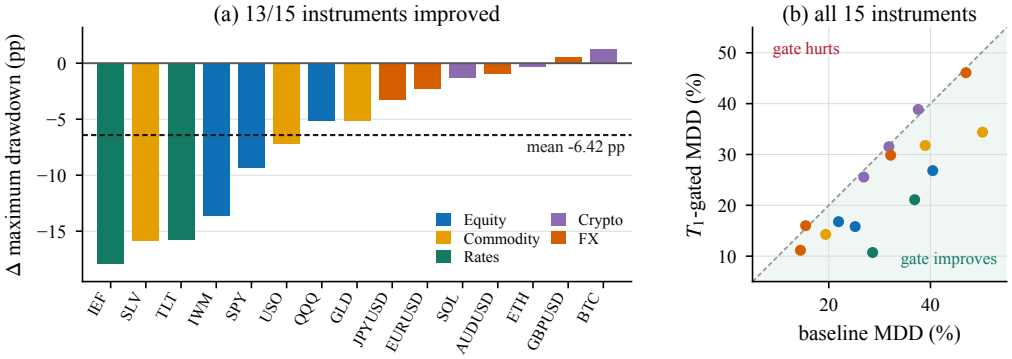


Figure 2: Main result, per instrument. (a) Change in maximum drawdown by asset class; 13 of 15 improve and the two that do not (BTC, GBPUSD) worsen by under 1.5 pp. (b) The same comparison as a scatter: points below the diagonal are improvements. Asset-class breakdown: Appendix Figure 18.

4 Experiments

4.1 Setup

Instruments. 15 liquid instruments across five asset classes: equities (SPY, QQQ, IWM), commodities (GLD, USO, SLV), rates (TLT, IEF), crypto (BTC, ETH, SOL) and FX (EURUSD, JPYUSD, GBPUSD, AUDUSD); daily OHLCV, 2014-2024 (SOL from 2020). The selection criterion was fixed before any result was examined: the most liquid, longest-history representative(s) of each asset class available via `yfinance` - not chosen or excluded on backtest performance. Except where stated, every quantity is computed independently per instrument.

Protocol, costs and metrics. A chronological 60/40 split for single-split numbers and 5-fold expanding-window walk-forward validation as the primary robustness check. The position at index t is applied to the return realised at index t; every input at index t is a function of information available no later than the close of day $t-1$ (Section 3.1). Costs are 10 bps per unit of one-way turnover (5-50 bps sweep and per-class costs below); metrics are annualised return (ARR), volatility (AVol), Sharpe ratio (ASR, no risk-free subtraction), maximum drawdown (MDD) and Calmar ratio (CR = ARR/MDD). Full hyperparameters: Appendix V.

4.2 Main result

The smooth T_1-gate reduces maximum drawdown on 13 of 15 instruments, mean reduction 6.42 pp (Table 1, Figure 2). Mean Sharpe moves from −0.007 to −0.029, both close to break-even; we quantify that small point-estimate cost formally next, where it does not survive. The hard variant trades a slightly better mean Sharpe for fewer instruments improved (Appendix G). Appendix Figure 6 shows the mechanism firing on SPY: T_1 crosses its train-calibrated threshold through the 2022 drawdown, the gate throttles exposure to near its floor for most of that year, and the two drawdown paths separate exactly there - the gate is not trading less on average so much as trading less *then* (cf. Corollary 1).

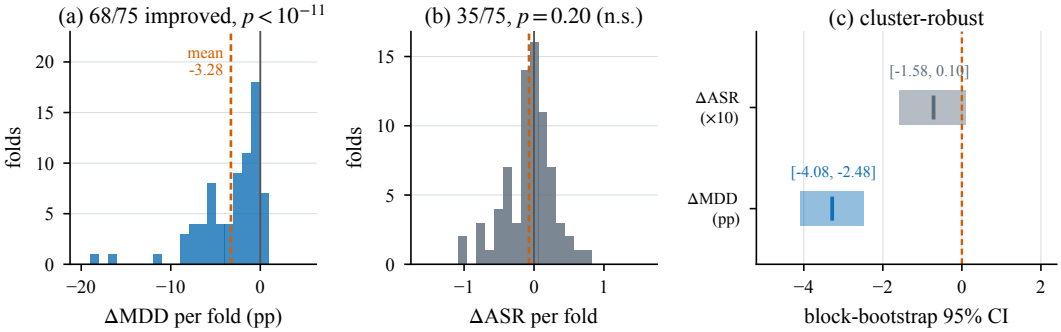


Figure 3: Walk-forward validation, 75 fold × instrument observations. The drawdown effect (a) is large and strongly significant; the Sharpe effect (b) is centred near zero and is not. A cluster-robust block bootstrap over instruments (c) confirms both readings.

4.3 Walk-forward validation and cluster-robust significance

A single split cannot separate a genuine effect from a fortunate split date. We re-run the smooth-gate comparison under 5-fold expanding-window walk-forward validation (75 fold × instrument observations). Condition-feature standardisation and gate parameters $(\mu, \sigma, z_{\text{thresh}})$ are both re-estimated independently within each fold from that fold's own training window only - no parameter is fit once and reused across folds.

The drawdown reduction holds on 68 of 75 folds, mean -3.28 pp, with a paired one-sided Wilcoxon signed-rank $p < 10^{-11}$ (Wilcoxon, 1945) (Figure 3). It is not an artefact of treating dependent expanding-window folds as independent: aggregating each instrument's folds to one mean first ($n = 15$, no pseudo-replication) still gives 15 of 15 instruments with a negative mean ΔMDD ($p = 3.1 \times 10^{-5}$), and a 20,000-draw block bootstrap over instruments (Künsch, 1989) gives a 95% CI of $[-4.08, -2.45]$ pp, excluding zero. Three independent tests agree.

The Sharpe point estimate is negative (mean $\Delta\text{ASR} = -0.072$) but under all three tests is not distinguishable from zero ($p = 0.203$ pooled; $p = 0.208$ per-instrument; bootstrap CI $[-0.158, +0.010]$). We therefore claim no Sharpe-ratio benefit anywhere in this paper. Per-instrument and per-fold tables: Appendix G.

4.4 Robustness: hyperparameters, direction signal, evaluation window, cost

Hyperparameters and direction signal. Sweeping T_1's own construction parameters one at a time ($k \in \{10, 20, 40\}$, $\tau \in \{0.25, 0.5, 1.0\}$, $w \in \{10, 20, 40\}$, gate fixed), 12-13 of 15 instruments improve at every one of the 9 settings; the gate's own parameters are similarly robust across a 15-point grid (11-15 of 15 everywhere). Replacing the traded direction with the classical 252-day-sign TSMOM signal (Moskowitz et al., 2012) makes the *baseline* much healthier (mean ASR 0.293 vs. -0.007 - notable on its own, since the simpler signal beats the richer MACD-combo one here) and the gate then improves **15 of 15** instruments (mean -6.38 pp); the benefit is not an artefact of one direction signal or parameter setting (Appendix J).

Evaluation window, transaction costs, and direction-feature entanglement. Over each non-crypto instrument's longest available history (from 2000, or real listing date if later) the gate improves 11 of 12 instruments (mean ΔMDD -3.33 pp; sole exception GLD). The gate raises mean daily turnover $1.52\times$ ($0.057 \rightarrow 0.087$) - possible because a time-varying multiplier is not a constant one (Appendix A) - and across a 5-50 bps grid the benefit is strongest at low cost (5 bps: 13/15, -6.77 pp) and weakens substantially without reversing sign at 50 bps (8/15, -1.29 pp); disclosed per-asset-class costs give 13/15 and -6.86 pp, if anything more favourable than the uniform headline. Dropping the direction signal from c_t entirely still improves 12 of 15 instruments (mean -4.66 pp) at essentially unchanged crisis AUROC ($0.757 \rightarrow 0.745$), so T_1's value is not an artefact of seeing the direction signal it gates. All three re-runs are reported in full, per instrument, in Appendix H and Figure 7.

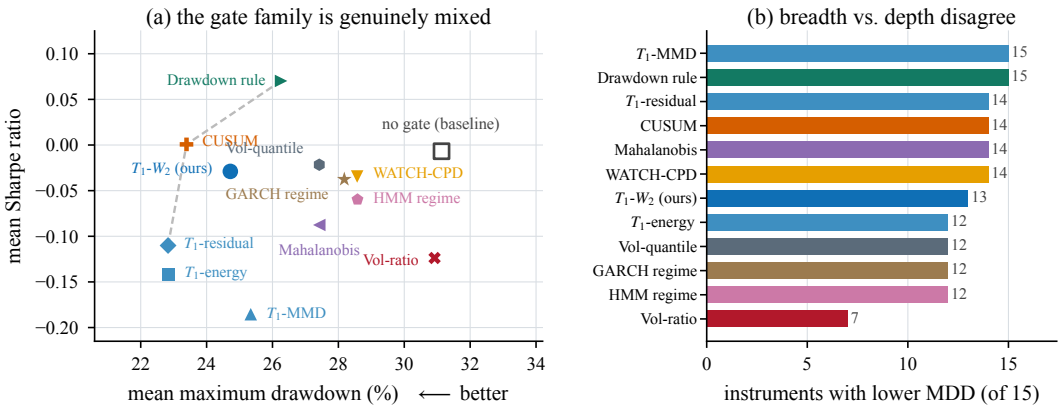


Figure 4: The competitor landscape. (a) Every gate uses the identical machinery of Equation 3; only the score changes. Eleven of twelve improve on the ungated baseline (open square) in drawdown, no gate dominates on both axes, and the Pareto frontier (dashed) contains three different constructions. (b) Improvement *breadth* and *depth* rank the family differently.

4.5 Is T_1 just re-deriving volatility, and how does it compare?

T_1 correlates strongly with same-day realized volatility on 14 of 15 instruments (Spearman $\rho \in [0.67, 0.93]$; exception SOL, $\rho = 0.30$; Appendix Figure 13). Since the baseline already volatility-scales, we test directly whether T_1's benefit is separable from volatility timing, and benchmark it against the fullest family of alternatives we could construct sharing identical gate machinery (Figure 4; summary table and per-instrument values in Appendix G, Table 7).

The naive competitor loses decisively. A gate on the raw short/long volatility ratio $\hat{\sigma}_5/\hat{\sigma}_{60}$, with identical machinery, improves only 7 of 15 instruments at mean MDD 30.92% - barely better than no gating at all. Despite T_1's strong volatility correlation, this is direct, quantitative evidence that it captures information a cheaper volatility-timing signal does not; we consider it the clearest single result in this paper's ablation family. The complementary test agrees: regressing T_1 on $(\hat{\sigma}_{20}, \hat{\sigma}_{60})$ train-only (mean $R^2 = 0.532$) and gating on the *residual* improves **14/15** instruments at mean MDD **22.83%**, better on both counts than the raw score. The non-volatility-explained component carries the benefit (Appendix N).

The fuller landscape is genuinely mixed, and we report it in full. Mahalanobis turbulence (Kritzman & Li, 2010) is a real competitor (14/15, smaller effect and worse ASR). Energy distance (Székely & Rizzo, 2013) is *better* on drawdown alone (22.85%, best in table) and substantially worse on Sharpe. An MMD variant (Gretton et al., 2012) improves **15 of 15** - with the drawdown rule, the only gates to do so - at the worst mean Sharpe tested. A two-parameter CUSUM statistic on $|r_t|$ (Page, 1954), with no feature space and no kNN at all, is the only gate with a non-negative mean Sharpe. Two conclusions follow and we hold both: T_1 is not uniquely optimal, and the *mechanism class* is broadly supported, since eleven of twelve gates here beat the ungated baseline on drawdown.

Two further probes, in the appendix. W_2 on order statistics discards within-window temporal order by construction - two windows that are permutations of one another receive distance exactly zero - yet both order-aware fixes we tested (cumulative-return W_2 and DTW (Sakoe & Chiba, 1978), each verified order-sensitive before use) *underperform* the default (Appendix M). And a single pre-specified feature-dimensionality sweep - fixed *before* seeing any result, precisely to avoid tuning until one configuration beat CUSUM on the data CUSUM's own number came from - narrows the gap to CUSUM without closing it, at the weakest mean ASR of every configuration tested (Appendix L).

4.6 Where the gate works, and where it does not

T_1's ability to discriminate crisis days is strongly heterogeneous across instruments (same-day AUROC 0.89-0.94 on most equities and FX, 0.50-0.52 on GLD and USO), and the gate's crisis-day exposure reduction tracks that heterogeneity closely and independently: the two separately-computed diagnostics correlate at $r = 0.98$ across instruments. T_1 also peaks *ahead* of the crisis label on most

instruments. The gate therefore works where it can see the crisis coming and is close to inert where it cannot - which is the behaviour we want, but it does mean the mean effect is carried by a subset of the universe. Full diagnostics, the lead/lag sweep, and the deliberately circular volatility-label control are in Appendix I and Figure 8.

4.7 Learned baselines and a negative result

Two learned-direction baselines (logistic regression and a small MLP, cross-entropy trained on T_1's condition features) train successfully but underperform fixed momentum on mean ASR (Table 1), a conclusion that survives a 5-seed study. Separately, following Lim et al. (2019) we attempted a single LSTM jointly predicting direction and magnitude via a turnover-penalised Sharpe-ratio loss, in six variants; all converge to the same qualitative failure - near-zero positions, mostly negative Sharpe. A warm-started sixth variant escapes that degenerate regime, which identifies the failure's mechanism as the initialisation rather than leaving it a guess, and still does not beat fixed momentum at this data scale. We report this as a negative result rather than omitting it. Full detail, per-instrument tables and the six variants: Appendix R and S, Figure 17(c).

4.8 Two tensions, made explicit

The two most interesting results in this paper are both negative, and both take the form of a frontier rather than a failure (Appendix Figure 15).

Calibration coverage versus drawdown benefit. What fraction of test days exceed the train-calibrated 90^{th}-percentile threshold? Mean 33.1% (median 33.3%, range 1.0-70.0%; Appendix P) - roughly three times nominal, direct evidence that exchangeability does not hold, exactly as the mechanism's own premise predicts. The obvious fix is an adaptive, ACI-style threshold (Gibbs & Candès, 2021), $\theta_{t+1} = \theta_t + \mathrm{lr}(\mathbb{1}[z_t > \theta_t] - \alpha)$, which restores nominal coverage (exceedance $33.1\% \to 11.3\%$ at $\mathrm{lr} = 0.10$) and erodes the drawdown benefit almost entirely (ΔMDD $-6.42 \to -1.40$ pp; improved $13/15 \to 11/15$). That coverage is restored is not luck: the recursion satisfies an exact identity.

Theorem 2 (Adaptive-threshold coverage identity)**.** *For any score sequence whatsoever - no exchangeability, no stationarity, no independence - and any* $\mathrm{lr} > 0$, $\left|\frac{1}{T}\sum_{t<T}\mathbb{1}[z_t > \theta_t] - \alpha\right| = |\theta_T - \theta_0|/(\mathrm{lr}\,T)$*, so empirical coverage converges to nominal at rate* $O(1/T)$ *whenever the thresholds stay bounded.*

The static threshold's $3\times$ exceedance is therefore not a calibration bug to be fixed - it *is* the mechanism: forcing nominal coverage throws away the shared regime-shift signal the gate was built to catch. Coverage and risk-management benefit are in genuine tension on this data, and we do not know how to have both.

Correlated gating versus portfolio benefit. Every result above is per instrument, and because T_1 triggers off shared macro information the gates could fire together during a common regime shift, eroding at the portfolio level the benefit each single-asset backtest shows in isolation. On an equal-weighted 15-instrument basket over a common 2022-onward window, portfolio MDD still falls substantially ($15.41\% \to 9.95\%$) - but the concern is real: mean pairwise gate correlation ($+0.307$) exceeds mean pairwise return correlation ($+0.157$), and portfolio Sharpe turns slightly negative. Extending the window to 2010-2024 without crypto (12 instruments, five real crises) reproduces both facts. A cross-sectional *relative* gate fully decorrelates ($+0.307 \to -0.045$) and nearly erases the benefit ($9.95\% \to 15.14\%$, barely below the ungated 15.41%); a softer peer-de-meaning variant traces the frontier between them, cutting gate correlation by two-thirds while keeping more than half the benefit. Some 'excess' correlation is genuine shared regime risk, not miscalibration, and a hard rank budget discards it along with the correlation. Full detail: Appendix O.

4.9 Multiple comparisons and the Deflated Sharpe Ratio

We report twelve gating signals and two hyperparameter sweeps without a formal multiple-testing correction, because the risk is asymmetric across the paper's two kinds of claim: the walk-forward test (Section 4.3) is a single, pre-specified primary test not selected from among the competing gates, while the competitor-landscape comparisons are framed throughout as descriptive robustness

checks. We additionally compute the Deflated Sharpe Ratio (Bailey & López de Prado, 2014) per instrument against an $N = 9$ trial family: mean 0.029, median 0.001, *0 of 15* instruments clear the conventional 0.95 bar. That confirms rather than contradicts the ASR-side non-significance already established; the MDD-side result, our primary claim, is not a Sharpe-based statistic and is not what DSR speaks to (Appendix Q).

5 LIMITATIONS

What it does not deliver, and where it is not the best choice. T_1-gating buys drawdown reduction at, at best, no Sharpe improvement; on point estimates it costs a little Sharpe, and under a Deflated Sharpe Ratio no instrument's gated Sharpe survives. It is a risk-management mechanism, not an alpha source. Nor is it uniformly dominant within its class: Mahalanobis turbulence, energy distance, an MMD variant, a drawdown rule and CUSUM are each competitive on at least one metric (Table 7), and a dedicated feature-space enrichment experiment narrowed but did not close the gap. We read the evidence as supporting the mechanism *class* rather than T_1's specific construction, and report the comparison in full because a selective account would misrepresent our own experiments.

Construction and scope. W_2 on order statistics discards within-window temporal order, and neither order-aware fix improves on it. Two to three instruments per asset class is modest for class-level generalisation claims, and the benefit is concentrated in rates, equities and commodities (Appendix U). Crisis-label sensitivity is tested against one alternative definition only. The benefit weakens substantially, though does not reverse, at high transaction costs, and even our class-specific costs remain bps-on-turnover models - this data source has no order-book or spread data to calibrate microstructure or slippage against. The primary analysis is single-asset, and the threshold is calibrated once rather than on a rolling basis; recalibrating deepens both the benefit and its Sharpe cost (Appendix T). GP-CPD (Saatçi et al., 2010) was scoped out rather than attempted - a documented decision, not a finding. Finally, although we verify the causal timing of every signal by perturbation test (Appendix B), readers reproducing this work should re-verify that chain independently rather than trust our audit alone.

What the formal verification covers. The Lean development (Appendix A) verifies the paper's mathematical statements. It does *not* verify that our Python implements them - that link is carried by the perturbation tests and unit checks of Appendix B, not by the proof assistant.

6 CONCLUSION

T_1-gated exposure reduces maximum drawdown on 13 of 15 single-asset instruments under walk-forward validation, cluster-robust re-analysis and three robustness sweeps, and it decisively beats the naive volatility-ratio competitor that T_1's own volatility correlation makes the obvious objection. Benchmarked against eleven alternative gates and a battery of further checks - several unfavourable, all reported - distribution-aware magnitude gating is a real, robust risk-management mechanism whose limits we can now state precisely: it does not buy Sharpe, it is not uniquely optimal in its class, and its benefit is in tension with both nominal coverage and portfolio decorrelation. Asking whether the market looks familiar is worth doing; keeping the answer useful once it must also be calibrated is the open problem.

AI USE STATEMENT

In this work, generative AI tools were used to (i) create or modify visual figures based on our experimental results, (ii) edit English grammar, phrasing, and vocabulary, and (iii) assist in producing Lean formalizations for checking selected mathematical equations. The tools were not used to generate data, formulate the research questions or theoretical claims, design or run experiments, analyze or interpret experimental results, or make scientific decisions. They were also not used to replace the authors' judgment about the scientific content.

We reviewed all AI-assisted work. The experimental data and results were checked against the original runs and stored result files; AI-assisted figures were checked against those results; and the Lean formalizations were compiled and checked by the authors to ensure that they corresponded to the equations stated in the paper. AI-assisted language edits were manually reviewed for meaning and accuracy. The authors take responsibility for the final content of this work, including all text, claims, figures, and formalizations produced with AI assistance.

REFERENCES

David H. Bailey and Marcos López de Prado. The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting and non-normality. *The Journal of Portfolio Management*, 40(5):94–107, 2014.

Jamil Baz, Nicolas Granger, Campbell R. Harvey, Nicolas Le Roux, and Sandy Rattray. Dissecting investment strategies in the cross section and time series. *SSRN Electronic Journal*, 2015. SSRN 2695101.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer, 2015.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

Leonardo de Moura and Sebastian Ullrich. The Lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28*, pp. 625–635. Springer, 2021.

Qianggang Ding, Haochen Shi, Jiadong Guo, and Bang Liu. TradExpert: Revolutionizing trading with mixture of expert LLMs. *arXiv preprint arXiv:2411.00782*, 2024.

Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

Kamil Faber, Roberto Corizzo, Bartlomiej Sniezynski, Michael Baron, and Nathalie Japkowicz. WATCH: Wasserstein change point detection for high-dimensional time series data. In *2021 IEEE International Conference on Big Data (Big Data)*, pp. 4450–4459, 2021. arXiv:2201.07125.

Fuli Feng, Huimin Chen, Xiangnan He, Ji Ding, Maosong Sun, and Tat-Seng Chua. Enhancing stock movement prediction with adversarial training. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5843–5849, 2019.

Siyu Gao, Yunbo Wang, and Xiaokang Yang. StockFormer: Learning hybrid trading machines with predictive coding. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4766–4774, 2023.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 1660–1672, 2021.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

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

Mark Kritzman and Yuanzhen Li. Skulls, financial turbulence, and risk management. *Financial Analysts Journal*, 66(5):30–41, 2010.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

Tong Li, Zhaoyang Liu, Yanyan Shen, Xue Wang, Haokun Chen, and Sen Huang. MASTER: Market-guided stock transformer for stock price forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 162–170, 2024.

Bryan Lim, Stefan Zohren, and Stephen Roberts. Enhancing time-series momentum strategies using deep neural networks. *The Journal of Financial Data Science*, 1(4):19–38, 2019. arXiv:1904.04912.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Alan Moreira and Tyler Muir. Volatility-managed portfolios. *The Journal of Finance*, 72(4):1611–1644, 2017.

Tobias J. Moskowitz, Yao Hua Ooi, and Lasse Heje Pedersen. Time series momentum. *Journal of Financial Economics*, 104(2):228–250, 2012.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

Long H.C. Pham. Decomposing out-of-distribution error in conditional flow matching via wasserstein geometry. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Yunus Saatçi, Ryan Turner, and Carl Edward Rasmussen. Gaussian process change point models. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 927–934, 2010.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, 2015.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations (ICLR)*, 2025.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

The mathlib Community. The Lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP)*, pp. 367–381, 2020.

Diego Vallarino. Adaptive market intelligence: A mixture of experts framework for volatility-sensitive stock forecasting. *arXiv preprint arXiv:2508.02686*, 2025.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2008.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Kieran Wood, Sven Giegerich, Stephen Roberts, and Stefan Zohren. Trading with the momentum transformer: An intelligent and interpretable architecture. *arXiv preprint arXiv:2112.08534*, 2021.

Kieran Wood, Stephen Roberts, and Stefan Zohren. Slow momentum with fast reversion: A trading strategy using deep learning and changepoint detection. *The Journal of Financial Data Science*, 4(1):111–129, 2022. arXiv:2105.13727.

Yiyao Zhang, Diksha Goel, Hussain Ahmad, and Claudia Szabo. RegimeFolio: A regime-aware ML system for sectoral portfolio optimization in dynamic markets. *arXiv preprint arXiv:2510.14986*, 2025.

APPENDIX

Roadmap. Appendix A gives the machine-checked mathematics and the proof of Proposition 1. Appendices B-D document the causal-timing verification protocol, the standardisation discipline, and one implementation detail of the distance computation. Appendix E completes the related-work discussion and Appendix F explains which literature we deliberately did not compare against numerically. Appendices G-H give the full per-instrument tables behind the main result and its robustness re-runs. Appendices I-L give the crisis diagnostics, the sensitivity grids, the extra baselines and the feature-space sweep. Appendices M-P give the sequence-aware and regime-model variants, the volatility decomposition, the portfolio experiments and the calibration coverage check. Appendices Q-T give the Deflated Sharpe Ratio, the learned baselines, the DMN reproduction and further extensions. Appendices V-W give hyperparameters and reproducibility details.

A MACHINE-CHECKED MATHEMATICS

Every mathematical statement in this paper is formalised and verified in Lean 4.33.1 against Mathlib (de Moura & Ullrich, 2021; The mathlib Community, 2020). The development is included as `lean/PaperMath.lean` in the supplementary material; it compiles with no `sorry`, and `#print axioms` reports that every theorem below depends only on Lean's three standard axioms (`propext`, `Classical.choice`, `Quot.sound`) - no additional assumptions are smuggled in.

What this does and does not buy. The formalisation verifies that our *mathematical claims* are true. It does not verify that our Python implements them: that link is carried by the perturbation tests and unit checks in Appendix B, not by the proof assistant. It does, however, pin down quantities the prose would otherwise leave approximate: the rank-to-z constant 28.87 used by the cross-sectional gate is $100/\sqrt{12} = 28.8675\ldots$ rounded, and the formalisation bounds that rounding at 2×10^{-4} relative (`rank_z_constant`) rather than leaving it asserted. The coverage identity likewise upgrades Section 4.8's adaptive-threshold result from an empirical observation to Theorem 2.

Table 2: The machine-checked statements, by paper location. Every entry compiles; names are the Lean identifiers in `lean/PaperMath.lean`.

Paper claim	Lean name	What is proved
Prop. 1, 1-D W_2 closed form	`sqCost_le_sqCost_comp_perm`	Monotone rearrangement minimises squared transport cost over all permutations
Thm. 1(i), gate range	`gate_mem_Icc`	$g(t) \in [g_{\min}, 1]$
Thm. 1(ii), monotonicity	`gate_strictAnti`, `gate_antitone`	g strictly decreasing in T_1 for $s > 0$
Thm. 1(iii), contraction	`gated_le_base`, `base_floor_le_gated`	$g_{\min} m \le g\, m \le m$ for $m \ge 0$
Thm. 1, threshold value	`gate_at_threshold`	$g = \frac{1+g_{\min}}{2}$ at $z = z_{\text{thresh}}$
Thm. 1, hard-gate limit	`gate_tendsto_gmin`	$g \to g_{\min}$ as $s \to \infty$ above threshold
§3.3, calibration in z-units	`zscore_threshold_iff`	Standardisation is order preserving, so calibrating in z-units is WLOG
Eq. 2, T_1 well-posed	`softmax_sum_one`, `t1_nonneg`, `t1_le_max`, `min_le_t1`	Softmax weights are a probability vector; T_1 is a convex combination of its neighbour distances
Cor. 1, timing effect	`sharpe_scale_invariant`	Sharpe is invariant under positive scaling
§4.1, MDD is a fraction	`drawdown_mem_Ico`	Drawdown $\in [0, 1)$ on a positive wealth path
§3.2, volatility targeting	`vol_target_exact`, `vol_target_mem_Icc`	$m_{\text{vol}}\hat{\sigma} = \sigma_{\text{target}}$ when unclipped; $m_{\text{vol}} \in [0, m_{\max}]$
Thm. 2, coverage identity	`aciThresh_telescope`, `aci_coverage_bound`	Exact telescoping identity and the $O(1/T)$ coverage gap, assumption-free
App. T, hybrid rules	`union_gate_eq_min`, `avg_gate_between`, `logSumExp_bounds`	Union rule = pointwise minimum of the two gates; average lies between them; LogSumExp is within $\log 2$ of the max
§4.5, MMD is a distance	`mmd_sq_eq_norm_sq`, `mmd_from_gram`	The estimator equals $\|\mu_a - \mu_b\|^2$, hence is non-negative
Fig. 12, order sensitivity	`raw_w2_order_blind`, `cumulative_w2_order_sensitive`	Raw W_2 gives 0 on a permutation; cumulative W_2 separates it
§4.5, energy distance	`energySq_symm`, `sqrt_clamp_eq_zero_iff`	Symmetry of the plug-in statistic; the clamp before the square root is safe
§Q.1, DSR	`dsrZ_strictAnti_in_SR0`	More trials strictly lower the DSR statistic
§4.4, turnover	`gating_can_increase_turnover`	An exposure-shrinking gate can still raise turnover
App. O.1, rank-to-z	`rank_z_constant`	$\lvert 100/\sqrt{12} - 28.87\rvert < 0.005$

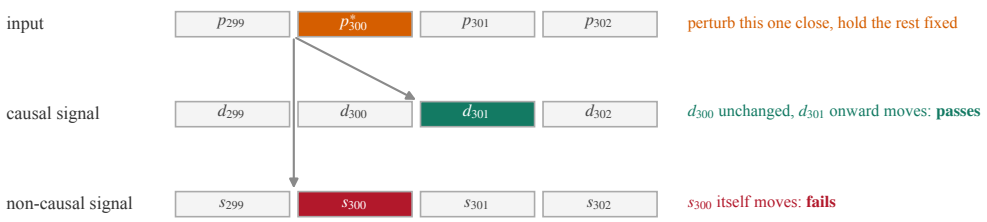


Figure 5: The causal-timing verification protocol. Perturbing a single input at index t^* and holding every other day fixed separates a signal that uses information only through t^*-1 (its value at t^* cannot move) from one that reads the current close (its value at t^* moves). Every signal used in this paper, including all eleven competing gates, passes this test.

A.1 Proof of Proposition 1

Proof. Write $\mathrm{cost}(f) = \sum_j (f_j - y_j)^2$. Expanding, $\mathrm{cost}(f) = \sum_j f_j^2 - 2\sum_j y_j f_j + \sum_j y_j^2$. Since σ is a bijection, $\sum_j x_{\sigma(j)}^2 = \sum_j x_j^2$, so the first and third terms are the same for x and $x \circ \sigma$, and the claim reduces to $\sum_j y_j x_{\sigma(j)} \le \sum_j y_j x_j$. Because x and y are both non-decreasing they *monovary*: if $y_i < y_j$ then $\neg(j \le i)$, hence $i < j$, hence $x_i \le x_j$. The rearrangement inequality for monovarying sequences then gives the required inequality. For the second statement, an optimal transport plan between two equally-weighted n-atom empirical measures on $\mathbb{R}$ under a convex cost is realised by a permutation, and by the above the identity coupling on the sorted atoms attains the minimum, so $W_2^2 = \frac{1}{n}\sum_j (x_{(j)} - y_{(j)})^2$ (Villani, 2008; Santambrogio, 2015). □

Both steps - the expansion and the rearrangement inequality - are the content of `sqCost_le_sqCost_comp_perm`. *Consequence for T_1.* Because our condition windows are single-asset (scalar) daily log returns, this closed form applies directly and exactly: no Monte-Carlo projection, no approximation error, and no dependence on a projection-count hyperparameter, unlike the general sliced-Wasserstein distance one would need for multi-dimensional (e.g. multi-asset) return vectors.

B Causal Timing: Convention and Verification

Everything in this paper rests on one convention: the position held at index t is applied to the return realised at index t, so every input to that position must be computable from information available no later than the close of day $t-1$. Rolling-window quantities satisfy this by construction - `realized_vol`, `vol_scaled_magnitude` and r_5 all use windows written as `[t-window:t]`, which exclude index t - but recursive quantities do not, and an exponentially-weighted moving average is the obvious trap: $\mathrm{out}[t] = \alpha x[t] + (1-\alpha)\,\mathrm{out}[t-1]$ incorporates $x[t]$ itself. The MACD-combo direction signal is built from EWMAs of the price, so its unshifted value at index t depends on p_t - the very close that determines the return the position earns. We therefore shift it by one day (`shifted[1:] = raw[:-1]`, Section 3.1), which is the definition of d_t used throughout.

The verification protocol. Rather than reason about each signal on paper, we test it. For a signal f and an index t^*, perturb the single input $x[t^*]$, hold every other day fixed, recompute f, and check that $f[t^*]$ is unchanged while $f[t^*+1]$ moves. A signal that passes cannot be using day t^*'s information at index t^*; a signal that fails is using information it should not have (Figure 5). Concretely, perturbing `prices[300]` leaves the shifted `macd_direction[300]` unchanged and changes index 301 onward, as required.

Applied to every signal, ours and the competitors'. The same test is run on every gating signal in Section 4.5, not only on T_1's own inputs, because a competitor that quietly sees the current close would make our comparison meaningless in our own favour's opposite direction. The CUSUM tail-risk gate consumes $|r_{t-1}|$ in its recursion; the drawdown-adaptive gate uses a drawdown computed as of the close of day $t-1$ (`rolling_drawdown_lag1`), while the contemporaneous `rolling_drawdown` is retained only for the crisis *label*, where same-day timing is what the label

means. The HMM and GARCH regime gates use the causal one-step-ahead *predicted* state probability and conditional variance respectively, never the filtered value. The vol-ratio, Mahalanobis, energy-distance, MMD, DTW, cumulative-W_2, WATCH-style and vol-quantile gates are causal by construction. Every one of them was perturbation-tested, and every one passes.

Why we report this at all. The convention above is easy to state and easy to get wrong, and a violation is invisible in aggregate performance statistics - it shows up as a better backtest, not as an error. Stating the protocol, rather than an assurance, is what lets a reader check the assumption instead of trusting it: the test is three lines of code and applies unchanged to any signal, ours or a reimplementation's.

C STANDARDISATION OF CONDITION FEATURES

Condition-feature z-scoring uses mean and standard deviation computed on $[0, \text{train_end})$ only, never on the full series. This matters even though the features are not forward-looking: standardising on the full series would leak test-period *scale* into the neighbour-search metric, changing which training days are retrieved as a query's nearest neighbours. In the walk-forward evaluation the statistics are re-fit independently within each fold, from that fold's own training window (Section 4.3); no standardisation parameter is fit once and reused across folds. The same discipline applies to the gate's calibration $(\mu, \sigma, z_{\text{thresh}})$ and to the train-only OLS coefficients of the volatility-residualised variant (Appendix N).

D WASSERSTEIN IMPLEMENTATION DETAIL: UNEQUAL WINDOW LENGTHS

Proposition 1 assumes equal sample sizes, which is what `compute_t1_series` always supplies in the paper's configuration: both windows have the fixed length w. The implementation nevertheless has to handle the truncated windows that occur near the start of a series, and does so by subsampling the longer window's order statistics onto $n = \min(|a|, |b|)$ indices via `np.linspace(0, len(a)-1, n).astype(int)`. We note explicitly that this is *nearest-index (truncating) selection*, not the linear interpolation of the empirical CDF that would be the textbook quantile-matching choice; the two agree at the endpoints and differ by at most one order-statistic gap in between. This path is never exercised for any number reported in this paper, but a variant with genuinely unequal windows would want interpolation instead.

E EXTENDED RELATED WORK

Where the T_1 score comes from. Pham (2026) decompose the out-of-distribution error of a conditional flow-matching generative model (Lipman et al., 2023) into three Wasserstein-geometric terms, of which the first - interpolation sparsity, the distance from a query condition to the training support - is the one we specialise here. In that setting the embedding space in which distances are measured is supplied by the model's own trained encoder. The observation this paper starts from is that for a single-asset trend follower the analogous space can simply be written down: four causal, interpretable summary statistics of the recent past, no encoder, no training. Everything else in T_1 - the kNN search, the softmax weighting, the distributional distance between trailing return windows - carries over unchanged, and the 1-D specialisation makes the distance exact rather than approximated (Proposition 1).

Classical volatility timing. Moreira & Muir (2017) show that scaling exposure down when realized volatility is high raises Sharpe ratios across a wide range of factors - the same mechanism as this paper's volatility-scaling baseline (Section 3.2), and the literature T_1-gating is layered on top of rather than competing with. Our volatility-ratio ablation (Section 4.5) is in effect a direct test of whether T_1 adds anything beyond a second, more convex application of the same volatility-timing idea.

Conformal calibration. Our gate calibrates its threshold on a training-period distribution of T_1 values, in the spirit of conformal risk control (Vovk et al., 2005; Gibbs & Candès, 2021), though we do not claim a formal coverage guarantee - and Section 4.8 shows directly why we cannot.

Mixture-of-experts and regime-aware trading. A substantial literature routes between sub-models conditioned on market regime, including mixture-of-experts time-series foundation models (Shi et al., 2025), volatility-regime-gated ensembles (Vallarino, 2025), LLM-based trading experts (Ding et al., 2024), jump-model regime allocation (Zhang et al., 2025), and transformer-based movement prediction (Feng et al., 2019; Gao et al., 2023; Li et al., 2024). These route or predict *direction* conditioned on regime; our contribution is orthogonal - we hold direction fixed and study only whether an OOD-motivated distance is a useful *magnitude* signal. Appendix F explains, method by method, why none of these is a controlled comparison for our task, and why we chose not to import their reported numbers.

F METHODS WE DELIBERATELY DID NOT COMPARE AGAINST NUMERICALLY

Every comparison in Section 4.5 shares this paper's data, instruments, direction signal, transaction-cost model and task formulation - the score inside Equation 3 is the only thing that changes. A second class of methods is frequently cited in this area but cannot be made into a controlled comparison without importing numbers produced under different conditions. We name them and say why, rather than either quietly omitting them or reporting cross-paper numbers as if they were comparable.

Table 3: Methods we name in Related Work but do not benchmark numerically, with the specific obstacle in each case.

Method	Venue	Why not a controlled comparison here
ALSTM (Feng et al., 2019)	IJCAI 2019	Next-day movement *classification* on fixed stock sets, not continuous position sizing: different loss, different label, different universe
StockFormer (Gao et al., 2023)	IJCAI 2023	Predictive-coding transformer on CSI300/CSI500; different market, different architecture scale, a faithful reproduction is weeks of work
MASTER (Li et al., 2024)	AAAI 2024	Cross-sectional stock *ranking* needing many correlated stocks per day; the task shape does not map onto single-asset sizing at all
Time-MoE (Shi et al., 2025)	ICLR 2025	A general forecasting foundation model, not a sizing rule; would require choosing a mapping from forecast to position that is our design decision, not theirs
LLM/jump-model regime allocation (Ding et al., 2024; Zhang et al., 2025)	2024-25	Route *direction* across sectors or experts; no single-asset magnitude rule to gate against
DMN, learned (Lim et al., 2019)	JFDS 2019	Attempted directly (Appendix S) and reported as a negative result; the original's numbers come from decades of data across 88 futures contracts and are not comparable to ours
GP-CPD (Saatçi et al., 2010)	ICML 2010	A faithful implementation is a substantially larger undertaking than the other additions here, and a rushed reimplementation would be worse than an honest omission

Reporting Section 4.5's family as the actual comparison - fully ours, controlled, reproducible - and citing the above as context is a stronger and more defensible position than a table of numbers copied from seven papers' seven different setups. Note that the classical volatility-scaling baseline we *do* use is DMN's own reported baseline in Lim et al. (2019), so the comparison is literature-grounded even where a learned-DMN row is absent.

G FULL PER-INSTRUMENT TABLES: MAIN RESULT

The hard-breaker variant. The hard gate $g_{\text{hard}} = \mathbb{1}[T_1 \leq \mu + z_{\text{thresh}}\sigma]$ has a slightly better mean Sharpe than the smooth gate (0.005 vs. -0.029) but improves fewer instruments (10/15 vs. 13/15) and a smaller mean drawdown effect (mean MDD 26.14% vs. 24.72%). We have no mechanistic explanation for the Sharpe difference beyond noting that both estimates are noisy around zero; per-instrument values are the right-hand columns of Table 4.

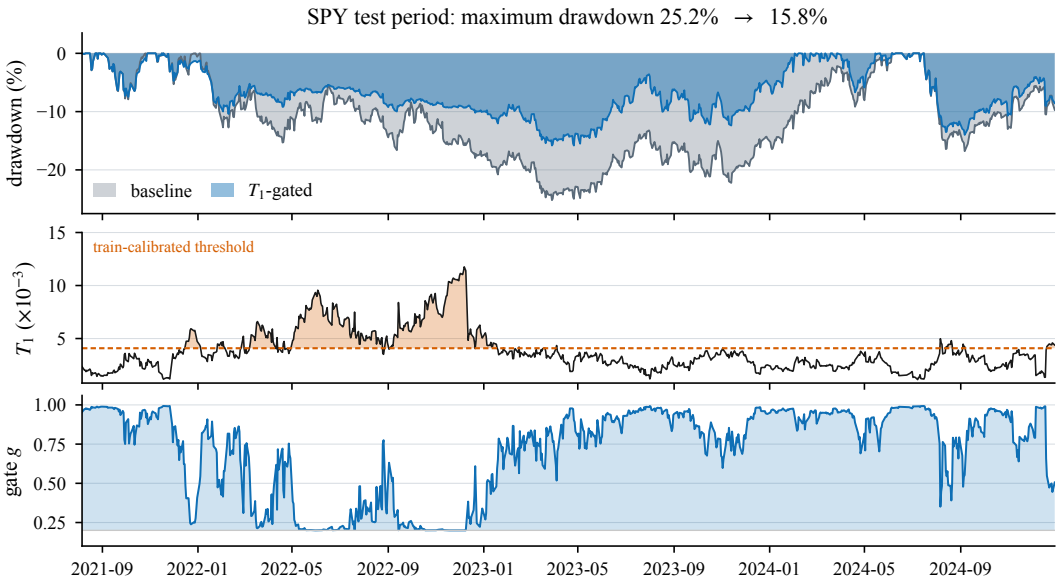


Figure 6: The mechanism in action on SPY's test period (Section 4.2). Top: drawdown paths, baseline versus gated. Middle: T_1 with its train-calibrated threshold, shaded where exceeded. Bottom: the resulting gate. The 2022 exceedance region is where the two drawdown paths separate - the gate is not trading less on average so much as trading less *then*, which is what Corollary 1 says any Sharpe difference must come from.

Table 4: Per-instrument annualised Sharpe ratio (ASR) and maximum drawdown (MDD, %), baseline versus both gate variants, single 60/40 split (Section 4.2).

Instrument	Class	Baseline		T_1-gate, smooth		T_1-gate, hard	
		ASR	MDD	ASR	MDD	ASR	MDD
SPY	Equity	0.220	25.19	0.429	15.82	0.704	14.50
QQQ	Equity	0.235	21.92	0.470	16.78	0.739	18.89
IWM	Equity	−0.680	40.45	−0.600	26.82	−0.523	25.75
GLD	Commodity	0.505	19.39	0.629	14.28	0.617	15.90
USO	Commodity	−0.584	38.96	−0.389	31.77	−0.470	36.17
SLV	Commodity	−1.019	50.24	−0.962	34.39	−0.864	35.59
TLT	Rates	−0.448	36.89	−0.472	21.11	−0.365	23.57
IEF	Rates	−0.107	28.62	−0.231	10.72	0.260	8.95
BTC	Crypto	0.062	37.59	−0.051	38.86	−0.071	38.90
ETH	Crypto	−0.318	31.82	−0.225	31.53	−0.202	32.85
SOL	Crypto	1.105	26.88	1.057	25.56	1.051	25.99
EURUSD	FX	0.180	32.17	−0.024	29.85	−0.259	31.76
JPYUSD	FX	0.964	14.42	0.562	11.16	0.171	17.83
GBPUSD	FX	0.549	15.45	0.448	16.00	0.429	15.45
AUDUSD	FX	−0.766	46.98	−1.076	46.08	−1.144	50.03
mean		*−0.007*	*31.13*	*−0.029*	*24.72*	*0.005*	*26.14*

Table 6 reports the fold-level minimum and maximum alongside the mean, which is the honest way to read the aggregate. Six of the fifteen instruments show a positive (worsened) ΔMDD in at least one fold, consistent with the 68/75 aggregate: the effect is strong and consistent in aggregate but not universal fold-by-fold. BTC (3/5 folds) is the weakest individual instrument on this measure; every other instrument improves in a majority of its folds, and 11 of 15 improve in at least 4 of 5.

H Full per-instrument tables: robustness re-runs

Choice of direction signal. Under the classical 252-day-sign TSMOM signal the baseline itself is far healthier (mean ASR 0.293 vs. −0.007) - notable on its own, since the simpler signal outperforms the richer MACD-combo one here - and the T_1-gate improves maximum drawdown on 15 of 15 instruments.

Table 5: The remaining backtest metrics, per instrument, single split: mean daily one-way turnover, annualised return (ARR, %), annualised volatility (AVol) and Calmar ratio (CR). The gate lowers realised volatility on every instrument while raising turnover on most (Section 4.4).

Instrument	Mean turnover			ARR (%)		AVol		CR	
	base	gate	ratio	base	gate	base	gate	base	gate
SPY	0.0426	0.0634	1.49	2.26	4.59	0.1587	0.1221	0.089	0.290
QQQ	0.0355	0.0539	1.52	2.50	4.77	0.1598	0.1125	0.114	0.284
IWM	0.0573	0.0748	1.31	−11.22	−6.39	0.1568	0.1015	−0.277	−0.238
GLD	0.0592	0.1259	2.13	7.00	7.08	0.1593	0.1203	0.361	0.496
USO	0.0356	0.0433	1.22	−10.09	−5.97	0.1600	0.1350	−0.259	−0.188
SLV	0.0494	0.0677	1.37	−15.84	−8.67	0.1570	0.0900	−0.315	−0.252
TLT	0.0758	0.1060	1.40	−7.84	−4.66	0.1554	0.0921	−0.212	−0.221
IEF	0.0983	0.1376	1.40	−2.57	−1.78	0.1448	0.0678	−0.090	−0.166
BTC	0.0233	0.0299	1.28	−0.29	−1.88	0.1608	0.1500	−0.008	−0.048
ETH	0.0203	0.0225	1.11	−6.34	−4.59	0.1639	0.1553	−0.199	−0.146
SOL	0.0167	0.0194	1.16	18.49	15.96	0.1659	0.1508	0.688	0.624
EURUSD	0.0844	0.1440	1.71	1.54	−0.91	0.1354	0.1135	0.048	−0.030
JPYUSD	0.0721	0.1251	1.74	14.40	5.08	0.1515	0.0965	0.999	0.456
GBPUSD	0.0752	0.1095	1.46	7.09	4.80	0.1434	0.1211	0.459	0.300
AUDUSD	0.1095	0.1787	1.63	−12.31	−12.48	0.1556	0.1174	−0.262	−0.271
mean	*0.0570*	*0.0868*	*1.52*	*−0.88*	*−0.34*	*0.1552*	*0.1164*	*0.076*	*0.059*

Table 6: Walk-forward validation, per instrument: 5 expanding-window folds each, with the fold-level minimum and maximum alongside the mean (Section 4.3).

Instrument	Folds impr.	ΔMDD (pp)			ΔASR		
		mean	min	max	mean	min	max
SPY	5/5	−5.79	−8.97	−1.59	0.105	−0.161	0.331
QQQ	5/5	−3.93	−7.85	−0.75	0.094	−0.384	0.343
IWM	5/5	−4.50	−8.56	−0.42	−0.405	−1.065	0.036
GLD	4/5	−3.59	−11.84	0.96	0.108	−0.339	0.729
USO	5/5	−3.21	−5.32	−1.09	0.171	−0.026	0.500
SLV	5/5	−4.56	−7.44	−0.82	−0.314	−0.810	0.120
TLT	4/5	−4.05	−16.78	0.75	−0.026	−0.793	0.672
IEF	5/5	−5.87	−18.73	−1.97	−0.272	−1.069	0.517
BTC	3/5	−0.38	−1.39	0.93	−0.037	−0.195	0.103
ETH	5/5	−2.34	−5.72	−0.12	0.055	−0.096	0.180
SOL	4/5	−1.40	−6.08	0.40	0.037	−0.147	0.289
EURUSD	4/5	−2.21	−5.06	0.03	−0.090	−0.350	0.318
JPYUSD	5/5	−3.16	−6.54	−0.06	−0.201	−0.538	0.166
GBPUSD	4/5	−0.70	−2.48	0.54	−0.125	−0.571	0.265
AUDUSD	5/5	−3.53	−8.08	−0.44	−0.174	−0.501	0.396
all folds	*68/75*	*−3.28*			*−0.072*		

Evaluation window. Over each non-crypto instrument's longest available history (from 2000-01-01, or real listing date if later - GLD from 2004, USO and SLV from 2006), the gate improves 11 of 12 available instruments; the sole exception, GLD, worsens slightly. The effect is weaker over this longer, more regime-diverse span than over 2014-2024 specifically, but remains directionally consistent, including across the 2008 financial crisis.

Class-specific transaction costs. The per-class assumptions (equities 3 bps, commodities 7, rates 4, crypto 20, FX 2) are disclosed, differentiated estimates, not a true microstructure calibration - this data source cannot support one. They give 13/15 improved at mean ΔMDD -6.86 pp and mean ΔASR $+0.049$, if anything more favourable than the uniform-10 bps headline.

Table 7: T_1-gate versus alternative gating signals, identical gate machinery, 15 instruments, single split. Baseline mean MDD 31.13%, mean ASR -0.007. Best per column in bold. The full 16-gate $\times$ 15-instrument matrix is Appendix Figure 10; per-instrument tables are Appendices K-H.

Gate	ASR	MDD	impr.	Gate	ASR	MDD	impr.
T_1, Wasserstein-2 (ours)	-0.029	24.72	13/15	CUSUM tail-risk	**0.001**	23.39	14/15
T_1, energy distance	-0.142	**22.85**	12/15	WATCH-style CPD	-0.034	28.57	14/15
T_1, MMD (median heur.)	-0.185	25.34	**15/15**	Mahalanobis turbulence	-0.088	27.42	14/15
T_1, cumulative W_2	-0.178	27.61	11/15	Drawdown rule	0.070	26.26	**15/15**
T_1, DTW	-0.083	25.27	13/15	Vol-quantile (level)	-0.022	27.42	12/15
T_1, vol-residualised	-0.110	22.83	14/15	Markov-switching HMM	-0.060	28.58	12/15
T_1, Mahalanobis kNN	-0.044	24.97	12/15	GARCH(1,1) regime	-0.038	28.18	12/15
T_1, no direction feature	-0.052	26.47	12/15	Vol-ratio (naive)	-0.124	30.92	7/15

Table 8: Maximum drawdown (%) for every competing gate, per instrument, single split (Section 4.5). All share the gate machinery of Equation 3; only the score changes. Columns: **VR** vol-ratio, **MAH** Mahalanobis turbulence, **ENE** energy distance, **MMD** maximum mean discrepancy, **WAT** WATCH-style CPD, **CUS** CUSUM, **VQ** volatility quantile, **DD** drawdown rule, **MkNN** Mahalanobis kNN search, **NoD** T_1 without the direction feature.

Instr.	Base	T_1	VR	MAH	ENE	MMD	WAT	CUS	VQ	DD	MkNN	NoD
SPY	25.2	15.8	26.7	21.1	12.9	19.0	22.2	13.3	22.0	15.5	17.6	16.0
QQQ	21.9	16.8	18.4	20.5	13.9	17.3	18.5	19.1	18.2	17.6	17.4	17.1
IWM	40.5	26.8	44.1	36.7	24.1	30.2	37.9	16.6	32.0	28.5	27.3	29.8
GLD	19.4	14.3	16.4	18.2	13.0	13.6	18.0	18.9	19.1	16.8	14.5	13.7
USO	39.0	31.8	37.1	33.0	26.4	24.6	35.0	32.0	36.1	38.5	30.9	34.7
SLV	50.2	34.4	44.0	45.8	35.9	48.4	46.0	34.8	44.7	43.0	35.2	40.4
TLT	36.9	21.1	40.0	29.0	15.8	31.3	28.7	13.1	22.1	22.2	20.9	25.0
IEF	28.6	10.7	30.3	16.5	14.7	23.5	20.0	6.3	12.3	16.7	10.1	16.2
BTC	37.6	38.9	41.4	34.5	38.3	31.5	37.5	36.7	37.7	33.8	39.2	38.5
ETH	31.8	31.5	29.3	30.2	27.3	27.2	31.7	34.2	31.6	31.5	31.6	30.9
SOL	26.9	25.6	25.4	21.4	17.1	11.7	26.8	26.1	26.8	26.7	25.1	26.7
EURUSD	32.2	29.9	32.9	33.2	28.9	30.4	32.4	32.1	31.1	30.8	30.1	30.0
JPYUSD	14.4	11.2	13.9	11.1	10.9	11.8	14.1	11.1	15.0	11.3	10.9	13.3
GBPUSD	15.4	16.0	15.9	14.4	16.5	13.4	15.2	15.3	15.3	15.4	15.8	16.2
AUDUSD	47.0	46.1	48.0	45.5	47.1	46.4	44.6	41.2	47.4	45.8	47.9	48.6
mean	*31.1*	*24.7*	*30.9*	*27.4*	*22.9*	*25.3*	*28.6*	*23.4*	*27.4*	*26.3*	*25.0*	*26.5*
improved	–	*13/15*	*7/15*	*14/15*	*12/15*	*15/15*	*14/15*	*14/15*	*12/15*	*15/15*	*12/15*	*12/15*

I CRISIS DIAGNOSTICS AND GATE SELECTIVITY

I.1 WHERE THE GATE WORKS: CRISIS DISCRIMINATION AND SELECTIVITY

T_1's same-day crisis AUROC (drawdown above a train-calibrated 80th percentile) is heterogeneous: strong on most equities and FX (0.89-0.94), weak on GLD/USO (0.50-0.52), intermediate on rates (0.69-0.74). The gate's crisis-day exposure reduction tracks this closely and independently (Appendix Figure 8(a); Pearson $r = 0.98$ between the two separately-computed diagnostics; e.g. ETH's mean gate falls $0.91 \to 0.37$ on crisis days, while GLD and USO show almost no selective drop). Sweeping the crisis label by $\pm$ trading days shows T_1 peaks *ahead* of the label on most instruments (Appendix Figure 8(b)). We deliberately avoid a volatility-based crisis label as *primary*, to avoid circularity with T_1's own volatility correlation; repeating the diagnostic with one anyway confirms the concern directly (mean AUROC jumps to 0.941, including 0.864/0.967 on GLD/USO) - the heterogeneous drawdown-based values are the economically meaningful ones.

We deliberately avoid a volatility-based crisis label as *primary*, because T_1's own strong volatility correlation would make such a diagnostic partly circular. Repeating the diagnostic with one anyway confirms the concern directly: mean AUROC jumps from 0.757 to 0.941, and the two instruments where the drawdown-based diagnostic is weakest are exactly the two that improve most (GLD $0.522 \to 0.864$, USO $0.497 \to 0.967$). The heterogeneous drawdown-based values are the economically meaningful ones.

Table 9: Three robustness re-runs, per instrument. *Left*: the classical 252-day-sign TSMOM direction signal replacing MACD-combo. *Middle*: each non-crypto instrument's longest available history (MDD %, start year shown). *Right*: per-asset-class transaction costs instead of a uniform 10 bps (Section 4.4).

Instr.	Classical TSMOM direction				Extended window (MDD)			Class-specific cost (MDD)		
	base ASR	base MDD	gate ASR	gate MDD	from	base	gate	bps	base	gate
SPY	0.495	25.1	0.452	21.1	2000	27.4	25.6	3	24.4	14.6
QQQ	0.483	23.5	0.484	18.7	2000	25.7	25.2	3	21.6	16.2
IWM	−0.097	27.8	−0.410	27.1	2000	45.2	41.3	3	39.0	23.9
GLD	−0.077	44.7	0.108	33.2	2004	24.5	25.0	7	18.8	13.6
USO	−0.244	36.8	−0.389	34.9	2006	39.0	33.2	7	38.5	31.1
SLV	−0.128	33.8	−0.251	25.4	2006	52.0	49.4	7	49.6	33.6
TLT	0.109	23.9	−0.026	15.8	2002	36.9	28.4	4	35.0	18.9
IEF	−0.014	32.2	0.250	9.2	2002	28.6	22.2	4	26.4	8.4
BTC	0.981	27.5	0.951	26.6	–	–	–	20	38.5	39.8
ETH	0.099	27.0	0.105	26.4	–	–	–	20	32.2	32.0
SOL	2.119	8.9	2.154	8.3	–	–	–	20	27.3	26.0
EURUSD	0.116	26.6	0.095	15.6	2003	42.5	40.0	2	28.4	25.5
JPYUSD	0.802	26.4	1.079	11.0	2000	47.0	41.2	2	13.6	10.0
GBPUSD	−0.103	30.4	−0.159	28.4	2003	23.3	21.4	2	13.8	14.2
AUDUSD	−0.140	27.9	−0.402	25.2	2006	50.7	49.9	2	42.8	39.1
mean	*0.293*	*28.2*	*0.269*	*21.8*		*36.9*	*33.6*		*30.0*	*23.1*
improved			*15/15*				*11/12*			*13/15*

Table 10: Transaction-cost sensitivity, per instrument (Section 4.4). Δ is gate minus baseline at each cost level.

Instrument	ΔMDD (pp) at cost				ΔASR at cost			
	5 bps	10 bps	20 bps	50 bps	5 bps	10 bps	20 bps	50 bps
SPY	−9.71	−9.37	−8.52	−6.43	0.240	0.209	0.146	−0.045
QQQ	−5.33	−5.14	−4.79	−3.94	0.268	0.236	0.171	−0.023
IWM	−14.69	−13.63	−11.63	−7.90	0.127	0.080	−0.014	−0.291
GLD	−5.21	−5.11	−4.85	3.94	0.209	0.124	−0.046	−0.552
USO	−7.46	−7.19	−6.66	−5.21	0.208	0.196	0.171	0.097
SLV	−16.08	−15.85	−15.23	−12.48	0.113	0.057	−0.053	−0.384
TLT	−16.00	−15.78	−13.36	−7.34	0.059	−0.024	−0.190	−0.665
IEF	−18.33	−17.90	−16.11	−1.88	0.047	−0.123	−0.460	−1.391
BTC	1.23	1.27	1.37	1.62	−0.107	−0.114	−0.128	−0.169
ETH	−0.33	−0.29	−0.22	0.00	0.096	0.093	0.087	0.071
SOL	−1.34	−1.32	−1.28	−1.16	−0.045	−0.048	−0.055	−0.076
EURUSD	−2.68	−2.32	−1.67	2.69	−0.123	−0.204	−0.363	−0.804
JPYUSD	−3.51	−3.26	−1.81	7.55	−0.298	−0.402	−0.606	−1.169
GBPUSD	0.47	0.55	0.72	4.48	−0.053	−0.101	−0.196	−0.464
AUDUSD	−2.55	−0.90	1.87	6.77	−0.209	−0.310	−0.509	−1.054
mean	*−6.77*	*−6.42*	*−5.48*	*−1.29*	*0.035*	*−0.022*	*−0.136*	*−0.461*
improved	*13/15*	*13/15*	*12/15*	*8/15*	*9/15*	*7/15*	*4/15*	*2/15*

J SENSITIVITY GRIDS

K EXTRA BASELINES AND ABLATIONS

Absolute volatility-quantile and drawdown-adaptive gates. Alongside the volatility-ratio gate, two further classical gates: one on the *level* of $\hat{\sigma}_{20}(t)$ (train-calibrated percentile, rather than a ratio of two horizons), and one directly on the instrument's own rolling drawdown as of the previous close (`rolling_drawdown_lag1`; the same quantity used for the crisis label in Appendix I, lagged one day so the gate never sees the close it is sizing into - see Appendix B). The vol-quantile gate improves drawdown on 12 of 15 instruments (mean ΔMDD -3.71 pp) - directionally useful, with a smaller average effect and worse mean ASR than T_1, Mahalanobis or CUSUM. The lag-1 drawdown-adaptive gate is the strongest of the three on breadth: 15/15 improved, mean ΔMDD -4.87 pp, mean ASR $+0.070$ - a real classical competitor on improvement breadth, though its mean drawdown effect remains smaller than T_1's (-6.42 pp).

Mahalanobis-versus-Euclidean kNN search. Distinct from the Mahalanobis-*gate* ablation (which replaces T_1's entire pipeline with a global Mahalanobis distance): here only the kNN *neighbour-*

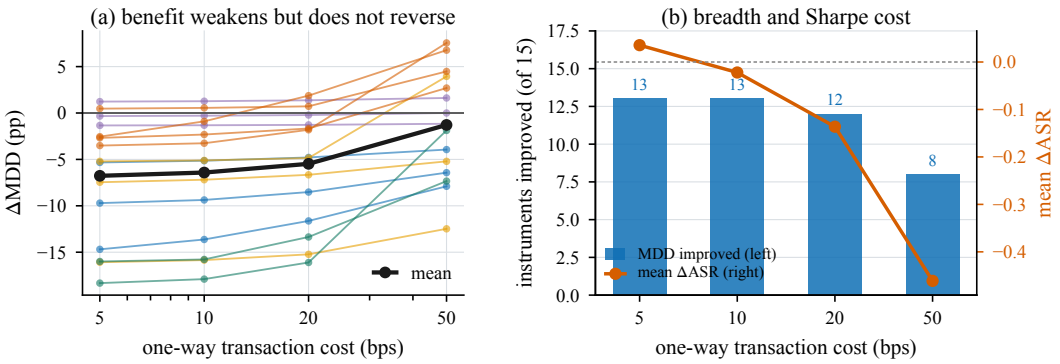


Figure 7: Transaction-cost sensitivity. (a) Per-instrument ΔMDD across the cost grid, with the cross-sectional mean in black; the benefit shrinks monotonically with cost but the mean never crosses zero. (b) The same sweep read as breadth (instruments improved) against the mean Sharpe cost.

Table 11: Crisis-detection and calibration diagnostics, all instruments, test period (Sections I.1 and 4.8). 'Peak lag' is the lead/lag at which crisis AUROC is maximised; negative means T_1 leads the label. SOL has no test-period crisis day at the calibrated threshold. Volatility correlations are Spearman ρ, all significant at $p < 10^{-4}$. Nominal exceedance is 0.10.

Instrument	Crisis %	AUROC	Peak lag	Mean gate			$\rho(T_1, \text{vol})$		Exceed.
				normal	crisis	drop	same-day	fwd-20d	rate
SPY	30.6	0.907	-4	0.845	0.450	0.395	0.927	0.522	0.292
QQQ	36.1	0.894	-4	0.780	0.346	0.435	0.890	0.552	0.420
IWM	28.3	0.720	-5	0.630	0.423	0.208	0.865	0.233	0.508
GLD	22.0	0.522	10	0.667	0.642	0.025	0.766	0.237	0.388
USO	12.4	0.497	10	0.825	0.880	−0.055	0.847	0.346	0.133
SLV	15.7	0.671	-2	0.470	0.318	0.152	0.744	0.485	0.700
TLT	62.6	0.690	3	0.664	0.478	0.186	0.829	0.349	0.556
IEF	58.4	0.735	-5	0.653	0.406	0.247	0.911	0.616	0.602
BTC	18.9	0.577	10	0.879	0.819	0.059	0.895	0.415	0.128
ETH	7.2	0.937	-1	0.914	0.372	0.542	0.680	0.279	0.137
SOL	0.0	n/a	None	0.930	n/a	n/a	0.295	0.193	0.010
EURUSD	16.6	0.907	-5	0.918	0.565	0.354	0.861	0.528	0.150
JPYUSD	22.0	0.734	10	0.743	0.449	0.294	0.911	0.612	0.402
GBPUSD	10.2	0.904	-5	0.873	0.444	0.428	0.790	0.510	0.201
AUDUSD	3.2	0.897	7	0.723	0.296	0.428	0.783	0.461	0.333

search metric changes, with Wasserstein-2 return-window scoring unchanged. Euclidean (default) improves 13/15 (mean ΔMDD −6.42 pp); Mahalanobis neighbour search improves 12/15 (mean −6.16 pp) - a small, not decision-relevant difference.

GP-CPD. Saatçi et al. (2010) combine Bayesian online change-point detection with Gaussian-process hazard modelling. A faithful implementation is a substantially larger undertaking than the other additions in this paper, and a rushed, possibly-buggy reimplementation would be worse than an honest omission; we did not attempt it. This is a documented scope decision, not a finding.

L Feature-space dimensionality sweep

The grid was fixed *before* seeing any result, precisely to avoid tuning until one configuration beat CUSUM on the data CUSUM's own number came from. We did not include volume or liquidity features, despite being a natural further candidate, because the volume data available for several FX pairs in our universe is unreliable or absent. The 2D (volatility-only) configuration is notable in its own right: it matches CUSUM's 14/15 improved-instrument count, ahead of the 4D default's 13/15, though with a smaller mean drawdown effect than either 4D or 6D. The 6D 'rich' space attains the best mean drawdown of the sweep at the worst mean Sharpe - plausibly curse-of-dimensionality on the Euclidean kNN search.

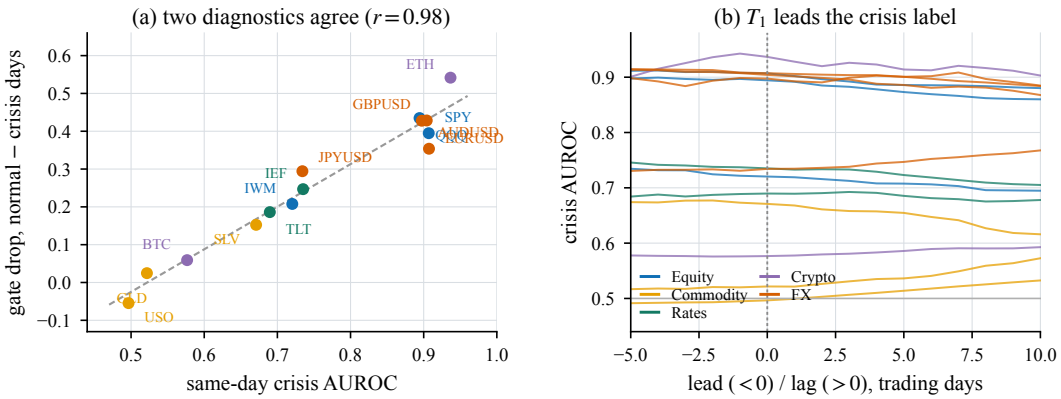


Figure 8: (a) Two independently-computed diagnostics agree: instruments where T_1 discriminates crisis days well are exactly the instruments where the gate selectively cuts exposure on those days ($r = 0.98$). GLD and USO sit at the bottom-left of both - T_1 cannot see their drawdowns coming, and the gate correspondingly does almost nothing. (b) Sweeping the crisis label forward and backward: on most instruments AUROC peaks at a negative lag, i.e. T_1 rises *before* the labelled crisis day.

Table 12: Crisis-label sensitivity: the drawdown-based label used throughout the paper versus a volatility-quantile regime label (Section I.1). The volatility label inflates AUROC on exactly the instruments where the drawdown label is weakest (GLD, USO), which is why we treat the drawdown-based values as the economically meaningful ones and the volatility-based ones as circular.

Instrument	Drawdown label		Volatility-regime label			
	crisis %	AUROC	regime %	AUROC	gate normal	gate crisis
SPY	30.6	0.907	27.9	0.968	0.862	0.366
QQQ	36.1	0.894	39.1	0.973	0.833	0.297
IWM	28.3	0.720	54.2	0.938	0.797	0.381
GLD	22.0	0.522	31.3	0.865	0.778	0.405
USO	12.4	0.497	13.8	0.966	0.904	0.386
SLV	15.7	0.671	56.9	0.857	0.625	0.311
TLT	62.6	0.690	56.8	0.917	0.792	0.362
IEF	58.4	0.735	62.4	0.960	0.844	0.308
BTC	18.9	0.577	11.6	0.945	0.915	0.506
ETH	7.2	0.937	4.8	0.966	0.903	0.320
SOL	0.0	n/a	0.0	n/a	0.930	n/a
EURUSD	16.6	0.907	12.6	0.978	0.929	0.379
JPYUSD	22.0	0.734	31.8	0.946	0.852	0.306
GBPUSD	10.2	0.904	16.9	0.969	0.924	0.364
AUDUSD	3.2	0.897	22.8	0.917	0.808	0.376
mean		*0.757*		*0.941*		

M SEQUENCE-AWARE DISTANCES AND REGIME MODELS

Cumulative-return W_2. Calibration (training-period percentile, smooth-gate mode) and neighbour search are identical to every other distance-metric ablation; only the inner distance function changes, to W_2 on cumulative rather than raw return windows. 11 of 15 instruments improve on MDD (mean -3.52 pp, mean MDD 27.61%), fewer than the default's 13/15, and the mean ASR is markedly worse (-0.178 vs. -0.029).

Dynamic time warping. Correctness was verified via three synthetic checks (`models/ood_t1.py`'s `_dtw_1d`) before use: non-zero on a permuted multiset where raw W_2 is exactly 0, exactly 0 on identical sequences, and finding a cheap warp for a time-shifted trend (5× cheaper than pointwise L_1). 13/15 improve on MDD (mean MDD 25.27%), matching the default's count, at a worse mean Sharpe (-0.083). The alignment cost is normalised by $(n+m)/2$, the standard near-equal-length path-length approximation, to keep the returned scalar on the same per-step scale as every other distance in this paper.

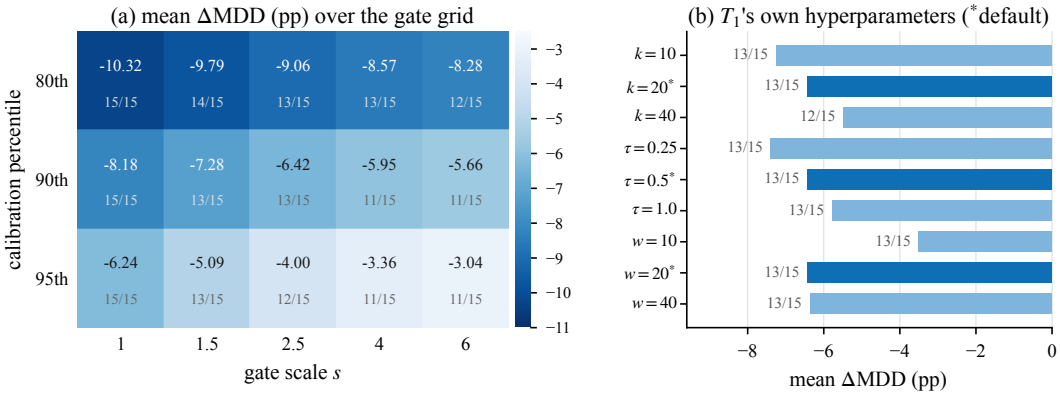


Figure 9: (a) The full 15-point gate-parameter grid; each cell gives the mean ΔMDD and the improved-instrument count. Every cell improves a majority of instruments, and the default (90th percentile, $s = 2.5$) is deliberately *not* the best cell - a lower percentile and gentler slope help more, at more turnover. (b) T_1's own construction hyperparameters, swept one at a time.

Table 13: Gate-parameter sensitivity, full 15-point grid over the calibration percentile and the sigmoid scale s, aggregated across the 15 instruments (Section 4.4). The default configuration is the 90th percentile at $s = 2.5$.

Percentile	Scale s	MDD improved	Mean ΔMDD (pp)	Mean ΔASR
80th	1	15/15	−10.32	−0.058
80th	1.5	14/15	−9.79	−0.060
80th	2.5	13/15	−9.06	−0.047
80th	4	13/15	−8.57	−0.026
80th	6	12/15	−8.28	−0.011
90th	1	15/15	−8.18	−0.045
90th	1.5	13/15	−7.28	−0.041
90th	2.5	13/15	−6.42	−0.022
90th	4	11/15	−5.95	−0.002
90th	6	11/15	−5.66	0.008
95th	1	15/15	−6.24	−0.036
95th	1.5	13/15	−5.09	−0.033
95th	2.5	12/15	−4.00	−0.029
95th	4	11/15	−3.36	−0.029
95th	6	11/15	−3.04	−0.031

Markov-switching HMM gate. A classical 2-state Gaussian HMM (Hamilton, 1989), fit by EM (scaled forward-backward) on train returns only for each instrument independently. The gate consumes the causal *one-step-ahead predicted* high-volatility state probability, never the filtered probability - verified by a mandatory perturbation test (`models/hmm_regime.py`) before any result here was trusted, mirroring Appendix B's methodology: perturbing return$[t]$ leaves predicted$[t]$ unchanged and only changes predicted$[t+1]$. 12/15 improve (mean MDD 28.58%, mean ASR -0.060). The fitted $\sigma_{\text{high}} > \sigma_{\text{low}}$ on every instrument, confirming the EM separates a genuine low/high-volatility regime pair rather than converging to a degenerate solution.

GARCH(1,1) regime gate. A hand-rolled, variance-targeting GARCH(1,1) (Bollerslev, 1986) fit on train data, using the causal one-step-ahead predicted conditional volatility (perturbation-verified the same way: perturbing r_{500} leaves predicted$[500]$ unchanged and only moves predicted$[501]$), gated with the identical sigmoid machinery. It improves 12/15 (mean MDD 28.18%, mean ΔMDD -2.95 pp, mean ASR -0.038) - comparable to but slightly weaker than the realized-vol quantile gate (12/15, 27.42%, -3.71 pp). Conditional volatility adds nothing over trailing realized volatility for this task.

N VOLATILITY DECOMPOSITION

Regressing T_1 on $(\hat{\sigma}_{20}, \hat{\sigma}_{60})$ by train-only OLS (mean $R^2 = 0.532$, range 0.231-0.726) and gating on the residual improves **14/15** instruments at mean MDD **22.83%** - better on both improved-

Table 14: T_1's own hyperparameter sensitivity: ΔMDD (pp) per instrument for a one-at-a-time sweep of neighbour count k, softmax temperature τ and return window w, gate parameters held at their defaults (* marks the default value; Section 4.4).

Instrument	k=10	k=20*	k=40	τ=0.25	τ=0.5*	τ=1.0	w=10	w=20*	w=40
SPY	−10.27	−9.37	−8.16	−10.52	−9.37	−8.72	−7.50	−9.37	−9.97
QQQ	−5.43	−5.14	−4.45	−5.76	−5.14	−5.01	−4.90	−5.14	−6.78
IWM	−13.73	−13.63	−12.05	−14.92	−13.63	−12.84	−8.28	−13.63	−14.63
GLD	−6.14	−5.11	−5.12	−5.79	−5.11	−4.67	−1.51	−5.11	1.12
USO	−8.34	−7.19	−6.50	−8.71	−7.19	−6.80	−4.90	−7.19	−5.25
SLV	−18.28	−15.85	−10.38	−19.62	−15.85	−12.77	−10.16	−15.85	−16.58
TLT	−17.07	−15.78	−13.32	−16.61	−15.78	−15.08	−10.51	−15.78	−17.69
IEF	−18.25	−17.90	−16.66	−18.81	−17.90	−16.02	−7.49	−17.90	−19.24
BTC	1.24	1.27	0.62	1.34	1.27	1.16	0.72	1.27	−0.30
ETH	−1.82	−0.29	0.02	−0.85	−0.29	−0.19	−0.22	−0.29	−3.62
SOL	−2.92	−1.32	−0.70	−1.99	−1.32	−0.92	−0.41	−1.32	−0.43
EURUSD	−4.09	−2.32	−1.30	−4.26	−2.32	−1.47	−0.63	−2.32	−0.27
JPYUSD	−2.40	−3.26	−3.66	−3.25	−3.26	−3.44	−0.32	−3.26	−4.15
GBPUSD	1.19	0.55	0.25	0.86	0.55	0.38	−0.35	0.55	−0.63
AUDUSD	−2.39	−0.90	−0.84	−2.00	−0.90	−0.24	3.99	−0.90	3.06
mean	*−7.25*	*−6.42*	*−5.48*	*−7.39*	*−6.42*	*−5.78*	*−3.50*	*−6.42*	*−6.36*
improved	*13/15*	*13/15*	*12/15*	*13/15*	*13/15*	*13/15*	*13/15*	*13/15*	*13/15*

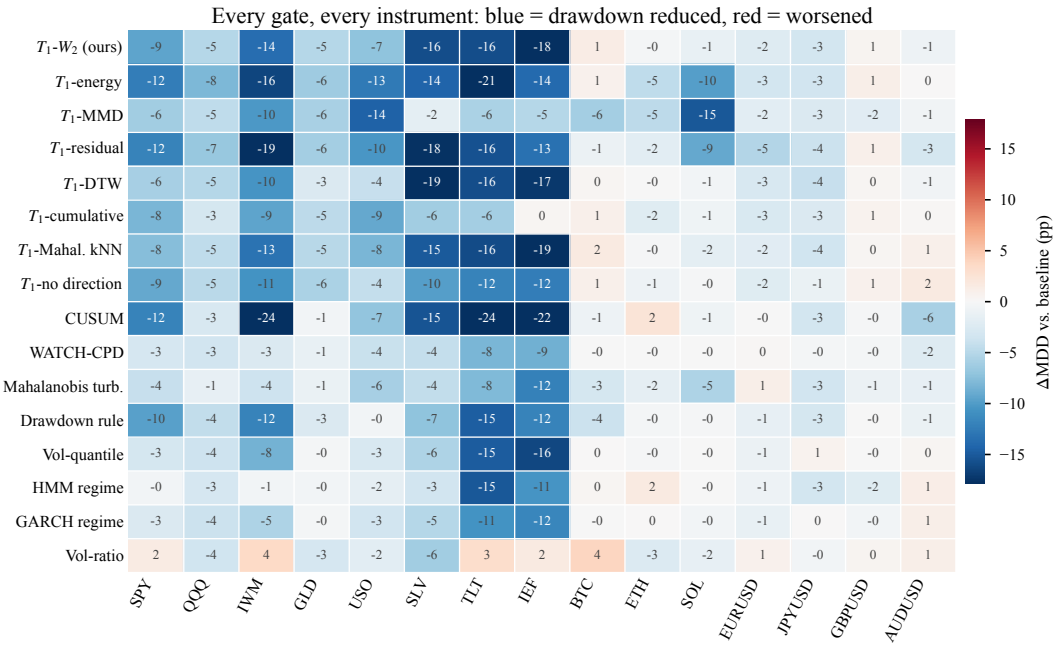


Figure 10: Every gate against every instrument: ΔMDD in percentage points relative to the ungated volatility-scaling baseline. Two structural facts are visible that no aggregate table shows. First, the benefit is concentrated in the same columns for *every* gate - IWM, SLV, TLT and IEF are where all of them work, and BTC, ETH, GBPUSD and AUDUSD are where none of them do. Second, the vol-ratio row is visibly different in kind from the rest, which is what Section 4.5's headline ablation measures.

count and mean MDD than the raw default (13/15, 24.72%) - at mean ASR -0.110. Regression coefficients are fit on train-period $(T_1, \hat{\sigma}_{20}, \hat{\sigma}_{60})$ only and applied unchanged to the full series, matching the train-only standardisation discipline of Appendix C. Per-instrument values are the 'vol-residualised' columns of Table 16. This is the strongest single piece of evidence that T_1's benefit is not volatility timing in disguise: removing the linearly volatility-explained component of the score does not weaken the gate, it strengthens it.

O PORTFOLIO EXPERIMENTS

Every single-asset result in this paper is computed independently per instrument. Since T_1 triggers off shared macro and volatility-regime information, gates could fire together across instruments

Table 15: Condition-feature dimensionality sweep, maximum drawdown (%) per instrument, single split (Section 4.5). The grid was fixed before any result was examined. The 6D 'rich' space adds the short/long volatility ratio and 20-day return skewness to the 4D default.

Instrument	Baseline	1D $[\hat{\sigma}_{20}]$	2D $+\hat{\sigma}_{60}$	3D $+r_5$	4D default	6D rich
SPY	25.2	17.1	14.7	16.0	15.8	14.7
QQQ	21.9	18.2	17.1	17.1	16.8	14.7
IWM	40.5	28.9	27.1	29.8	26.8	23.1
GLD	19.4	14.7	13.5	13.7	14.3	14.1
USO	39.0	36.2	31.7	34.7	31.8	29.5
SLV	50.2	43.9	40.6	40.4	34.4	31.0
TLT	36.9	25.4	24.8	25.0	21.1	21.9
IEF	28.6	18.3	13.7	16.2	10.7	11.1
BTC	37.6	36.7	36.6	38.5	38.9	39.7
ETH	31.8	30.8	30.9	30.9	31.5	28.8
SOL	26.9	26.7	26.4	26.7	25.6	23.3
EURUSD	32.2	31.2	29.8	30.0	29.9	27.2
JPYUSD	14.4	14.0	10.2	13.3	11.2	10.2
GBPUSD	15.4	15.6	16.1	16.2	16.0	15.8
AUDUSD	47.0	47.2	45.7	48.6	46.1	46.0
mean	*31.13*	*27.00*	*25.26*	*26.47*	*24.72*	*23.40*
mean ASR	*−0.007*	*−0.022*	*−0.062*	*−0.052*	*−0.029*	*−0.127*
improved	–	*13/15*	*14/15*	*12/15*	*13/15*	*13/15*

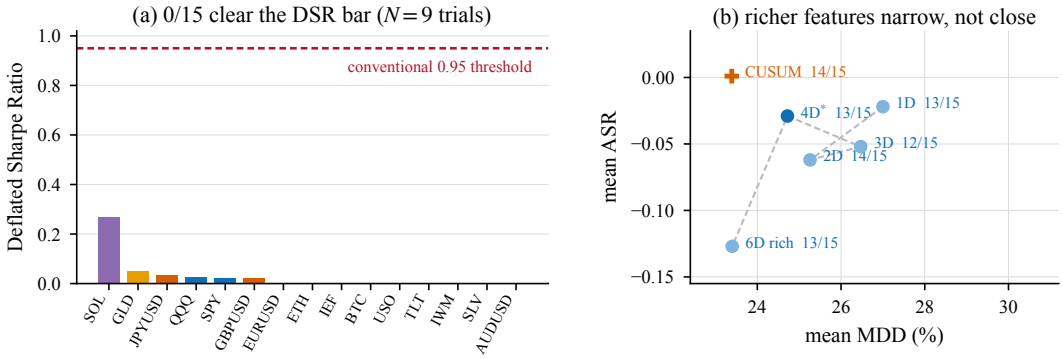


Figure 11: (a) Deflated Sharpe Ratio per instrument against the $N = 9$ trial family: no instrument clears the conventional 0.95 bar, which is formal confirmation of the ASR-side non-significance rather than a contradiction of the drawdown claim. (b) The pre-specified feature-space sweep against CUSUM: richer feature spaces move towards CUSUM's drawdown level but at a monotonically worse Sharpe, and do not reach it.

during a common regime shift, eroding at the portfolio level the diversification benefit each single-asset backtest shows in isolation. We test this directly rather than leaving it as an untested limitation.

Design. An equal-weighted basket over a common calendar test window (2022-01-01 onward, the latest date every instrument has a full pre-2022 training history for), with per-instrument T_1 and gate mechanics unchanged and no cross-asset information entering T_1 itself. Non-trading days contribute 0 to that day's portfolio return, which handles the fact that equities close five days a week while crypto trades seven. This common-window choice is the one design decision this experiment adds beyond the rest of the paper's per-instrument 60/40 split, and it is needed only to align dates for aggregation.

Reading the result. Portfolio MDD falls substantially ($15.41\% \rightarrow 9.95\%$), so the single-asset drawdown benefit does not cancel out once assembled into a basket - but the correlated-gating concern is real, not hypothetical: mean pairwise gate correlation ($+0.307$) exceeds mean pairwise return correlation ($+0.157$) by $+0.150$, and portfolio Sharpe turns slightly negative, matching the single-asset pattern.

Window length. Dropping crypto extends the common window to 2010-2024 (12 non-crypto instruments, 15 years, five real crises; common start 2010-01-01, chosen because SLV, the latest-listed of the twelve at 2006-04-28, needs over three years of pre-test training history under this paper's

Table 16: Sequence-aware and regime-model gate variants, per instrument, single split (Sections 4.5). Baseline and default-T_1 MDD (%) shown for reference.

Instr.	Base	T_1	Cumulative W_2		DTW		Vol-residualised		HMM regime		GARCH regime	
	MDD	MDD	ASR	MDD	ASR	MDD	ASR	MDD	ASR	MDD	ASR	MDD
SPY	25.2	15.8	0.191	17.1	0.336	19.2	0.451	13.3	0.195	24.8	0.230	22.6
QQQ	21.9	16.8	0.064	18.8	0.573	16.8	0.296	15.2	0.197	18.9	0.349	18.1
IWM	40.5	26.8	−0.627	31.0	−0.744	30.0	−0.519	21.1	−0.833	39.9	−0.755	35.0
GLD	19.4	14.3	0.269	14.7	0.463	16.9	0.601	13.3	0.508	18.9	0.542	19.0
USO	39.0	31.8	−0.318	29.9	−0.474	34.9	−0.345	28.6	−0.602	37.4	−0.548	35.6
SLV	50.2	34.4	−1.046	44.0	−1.015	31.1	−0.944	32.5	−1.083	46.9	−1.024	45.2
TLT	36.9	21.1	−0.785	30.8	−0.632	21.1	−0.481	21.4	−0.374	21.7	−0.539	26.2
IEF	28.6	10.7	−0.902	29.0	−0.466	11.4	−0.321	15.3	−0.288	18.1	−0.317	16.9
BTC	37.6	38.9	−0.073	38.5	−0.023	38.0	−0.076	36.5	−0.015	38.0	0.022	37.5
ETH	31.8	31.5	−0.249	29.7	−0.247	31.7	−0.265	29.9	−0.289	33.5	−0.266	31.8
SOL	26.9	25.6	0.999	26.2	1.098	26.4	0.668	17.8	1.153	26.9	1.118	26.4
EURUSD	32.2	29.9	0.089	29.4	−0.009	29.3	−0.205	26.9	0.041	31.2	0.175	30.8
JPYUSD	14.4	11.2	0.302	11.8	0.689	10.3	0.149	10.6	0.954	11.1	0.874	14.5
GBPUSD	15.4	16.0	0.377	16.0	0.324	15.8	0.452	16.3	0.493	13.3	0.520	15.0
AUDUSD	47.0	46.1	−0.961	47.2	−1.115	46.4	−1.114	43.8	−0.952	48.2	−0.945	48.0
mean	*31.1*	*24.7*	*−0.178*	*27.6*	*−0.083*	*25.3*	*−0.110*	*22.8*	*−0.060*	*28.6*	*−0.038*	*28.2*

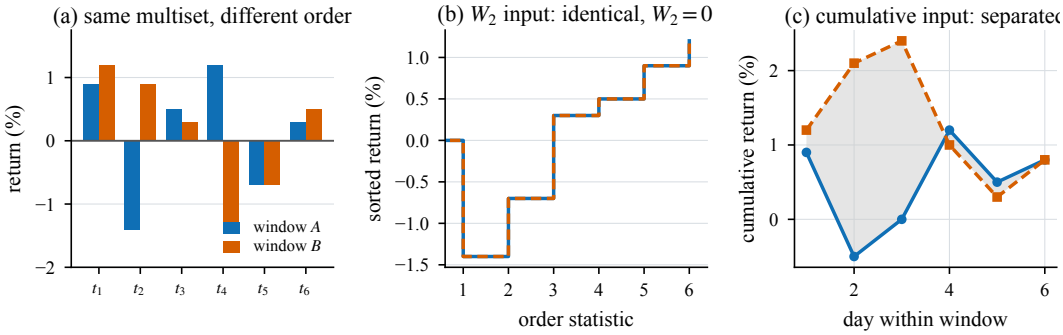


Figure 12: Why the two order-aware variants exist, and exactly what they change. (a) Two 20-day windows (shown at length 6 for legibility) containing the same returns in a different order. (b) The default score sees only the sorted order statistics, which are identical, so it assigns distance exactly zero. (c) The cumulative variant sees two genuinely different paths. Both statements are machine-checked (Appendix A). Empirically, the order-aware variants nevertheless *underperform* the default (Table 16).

MIN_TRAIN_DAYS $= 300$ threshold). The same pattern holds: MDD improves ($15.33\% \rightarrow 13.44\%$) and gate correlation ($+0.313$) again exceeds return correlation ($+0.101$). Neither fact is an artefact of a short window.

O.1 Portfolio-aware gate variants

Each instrument's T_1 is standardised on its own train-period (μ, σ), as in every other gate in this paper. On each test day with at least 8 co-trading instruments, the *cross-sectional* gate ranks each instrument's standardised z-score among currently-active peers, converts the rank to a percentile, and re-expresses it in z-units via the continuous-uniform approximation $z_{\text{rank}} = (\text{pct} - 50)/28.87$ - where 28.87 is $100/\sqrt{12}$, the standard deviation of a $\text{Uniform}(0, 100)$ percentile, rounded (machine-checked bound: Appendix A) - before applying the identical sigmoid gate shape of Equation 3. On the 31.5% of days with fewer than 8 co-trading instruments (weekends and holidays, when only crypto trades) the gate falls back to that instrument's own independent threshold: a disclosed hybrid, not a hidden one.

The hard rank gate fully decorrelates ($+0.307 \rightarrow -0.045$) and nearly erases the benefit (MDD $9.95\% \rightarrow 15.14\%$, barely below the ungated 15.41%). A *soft* variant, $z_{\text{soft}} = z_{\text{own}} - \alpha \cdot \text{mean}(z_{\text{active peers}})$, keeping per-instrument standardisation and the sigmoid threshold unchanged, traces the frontier between the two: $\alpha = 1$ decorrelates as completely as the rank (-0.025) but still erases most of the benefit, while $\alpha = 0.5$ cuts gate correlation by two-thirds ($+0.307 \rightarrow +0.105$) while retaining more than half the drawdown benefit ($9.95\% \rightarrow 12.47\%$ against the 15.41% ungated baseline). The decorrelation/benefit trade-off is a continuous frontier (Figure 15(b)), not a binary

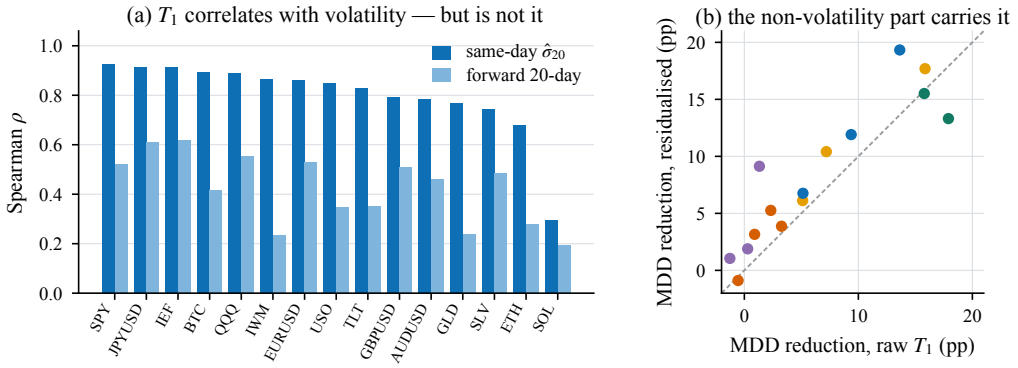


Figure 13: (a) T_1 against realized volatility, per instrument: the same-day correlation is high almost everywhere, which is what makes the volatility objection the sharpest one available; the forward-20-day correlation is much weaker, so T_1 is not simply a volatility forecast either. (b) Drawdown reduction from the raw score against the reduction from the volatility-residualised score: most points lie above the diagonal, i.e. the non-volatility component alone does at least as well.

Table 17: Portfolio backtests. *Top*: the 15-instrument basket, 2022-2024, with the independent gate and two portfolio-aware variants. *Bottom*: the 12-instrument non-crypto basket over the extended 2010-2024 window. Mean pairwise *return* correlation is $+0.157$ (top block) and $+0.101$ (bottom block).

Portfolio	ARR (%)	ASR	MDD (%)	CR	Mean gate corr.
15 instruments, 2022-01-06 to 2024-12-30					
Baseline (vol-scaled, no gate)	-0.00	0.027	15.41	-0.000	-
T_1-gated, independent threshold	-0.20	-0.032	**9.95**	-0.020	$+0.307$
Cross-sectional rank gate	-0.76	-0.138	15.14	-0.050	-0.045
Soft gate, $\alpha = 1$ (peer de-mean)	-0.70	-0.126	14.46	-0.048	-0.025
Soft gate, $\alpha = 0.5$ (partial blend)	-0.45	-0.081	12.47	-0.036	$+0.105$
12 non-crypto instruments, 2010-01-08 to 2024-12-30					
Baseline (vol-scaled, no gate)	0.61	0.130	15.33	0.040	-
T_1-gated, independent threshold	0.23	0.070	**13.44**	0.017	$+0.313$

choice - some 'excess' correlation is genuine shared regime risk, not miscalibration, and a hard rank budget throws that signal away along with the correlation.

P Calibration coverage

What fraction of test days exceed the train-calibrated 90^{th}-percentile threshold? The per-instrument exceedance rates are the last column of Table 11 and Figure 16(c): mean 0.331, median 0.333, range 0.010-0.700, with only 4 of 15 instruments falling inside even a generous $2\times$ band around the nominal 0.10 (i.e. 0.05-0.20). This is a direct, quantitative rejection of the naive train/test exchangeability assumption a formal conformal guarantee would need (Section 3.2's Remark) - and it is exactly what the mechanism's own premise, that regimes shift, predicts. The calibration is usefully heuristic, not a formal guarantee, and we do not claim otherwise anywhere in this paper.

Q Deflated Sharpe Ratio, full table

Q.1 Multiple comparisons and the Deflated Sharpe Ratio

This paper reports twelve gating signals and two hyperparameter sweeps without a formal multiple-testing correction. We consider the risk asymmetric across the paper's two kinds of claims: the walk-forward test (Section 4.3) is a single, pre-specified primary test of T_1-gate versus baseline, not selected from among the competing gates, so it is not directly subject to this concern; the competitor-landscape comparisons are framed throughout as descriptive robustness checks rather than repeated hypothesis tests of one claim. We additionally compute the Deflated Sharpe Ratio (Bailey & López de Prado, 2014) per instrument against an $N = 9$ gate trial family: mean DSR

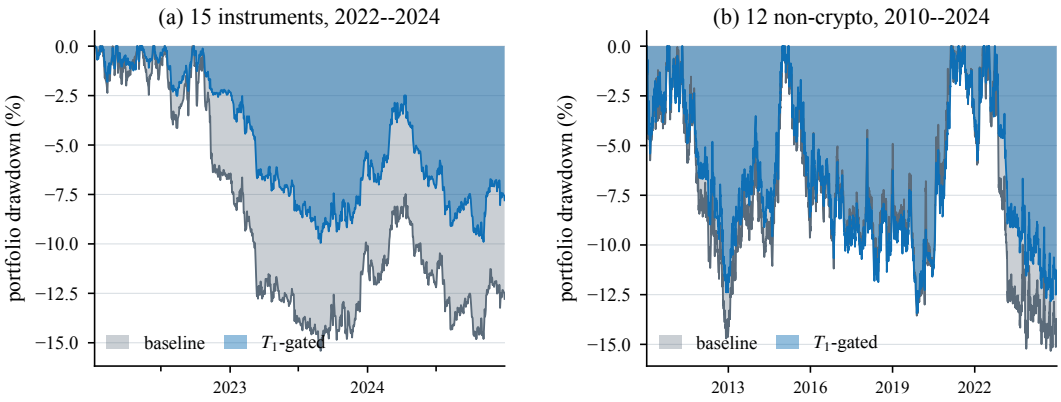


Figure 14: Portfolio drawdown paths, baseline versus T_1-gated. (a) The 15-instrument basket over the common 2022-2024 window. (b) The 12-instrument non-crypto basket over 2010-2024, which spans five real crises. The drawdown benefit survives aggregation in both, which is the question the experiment was built to answer.

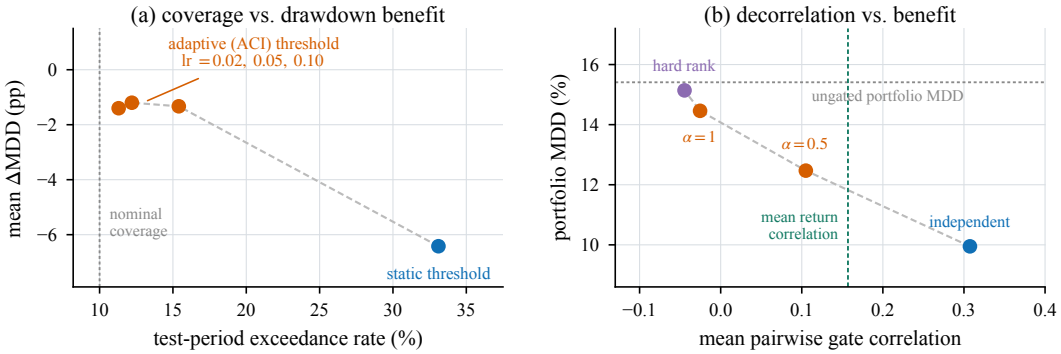


Figure 15: Both negative results are frontiers, not walls. (a) Restoring nominal calibration coverage removes almost all of the drawdown benefit - the static threshold's over-exceedance is carrying the signal. (b) Decorrelating the gate across a portfolio removes the benefit with the correlation, but the soft variant traces a continuous trade-off rather than a binary choice.

0.029, median 0.001, and *0 of 15* instruments exceed the conventional 0.95 threshold (Appendix Figure 11(a)). That is formal confirmation, not contradiction, of the ASR-side non-significance already established above - and the MDD-side result, this paper's primary claim, is not a Sharpe-based statistic and is not what DSR speaks to. More trials can only lower the DSR (Appendix A, `dsrZ_strictAnti_in_SR0`), which is exactly why we report the full family rather than a selected subset.

SR_0 - the expected maximum Sharpe ratio achievable by chance among $N = 9$ trials - is computed from the empirical variance of the nine gates' mean ASR values in Table 7 following Bailey & López de Prado (2014) exactly, giving $\mathrm{SR}_0 = 0.0958$ (daily). The DSR is applied per instrument to that instrument's own daily net-return series, which is its intended use case, and is *not* pooled across instruments: pooling would treat autocorrelated daily observations as independent, which is the same pseudo-replication error that Section 4.3's cluster-robust re-analysis exists to avoid.

R Learned Direction Baselines

R.1 Learned Baselines and a Negative Result

A logistic regression and a small MLP, cross-entropy trained on T_1's condition features to predict the 5-day return sign, both train successfully but underperform fixed momentum on mean ASR (-0.534, -0.253 vs. -0.007; Table 1), though each beats it more often per instrument (4/15 and 6/15) than the mean comparison alone suggests. A 5-seed study confirms the sign and the seed sensitivity (Appendix R.1).

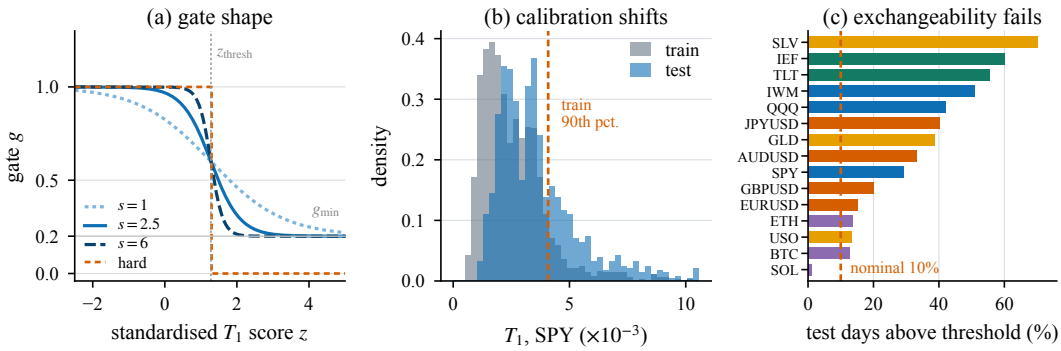


Figure 16: (a) The gate of Equation 3 at three steepnesses, with the hard variant as the $s \to \infty$ limit (Theorem 1). (b) T_1's train and test distributions on SPY against the train-calibrated threshold: the test distribution has visibly shifted right. (c) The consequence, per instrument - the fraction of test days above the train-calibrated 90^{th} percentile is roughly three times nominal on average, a direct rejection of train/test exchangeability and exactly what the mechanism's own premise predicts.

Table 18: Deflated Sharpe Ratio (Bailey & López de Prado, 2014) per instrument, computed on that instrument's own daily net-return series against the $N = 9$ gate trial family ($\text{SR}_0 = 0.0958$, from the empirical variance of the nine gates' mean ASR). Applied per instrument, never pooled: pooling would treat autocorrelated daily observations as independent, the same pseudo-replication error the cluster-robust re-analysis of Section 4.3 exists to avoid.

Instrument	$\widehat{SR}$ (daily)	n	Skew γ_3	Kurt. γ_4	DSR
SPY	0.0270	855	−0.514	5.79	0.0230
QQQ	0.0296	855	−0.371	5.99	0.0273
IWM	−0.0378	855	0.151	7.22	0.0001
GLD	0.0396	855	−0.213	5.90	0.0511
USO	−0.0245	855	−0.476	8.29	0.0002
SLV	−0.0606	855	−0.041	8.44	0.0000
TLT	−0.0297	855	0.380	7.74	0.0001
IEF	−0.0145	855	0.245	9.19	0.0006
BTC	−0.0032	1252	0.259	8.43	0.0002
ETH	−0.0142	792	0.057	9.13	0.0010
SOL	0.0666	439	0.532	4.66	0.2672
EURUSD	−0.0015	895	0.356	5.70	0.0018
JPYUSD	0.0354	895	0.243	7.95	0.0350
GBPUSD	0.0282	895	0.275	5.63	0.0213
AUDUSD	−0.0678	895	−0.017	5.21	0.0000
mean					*0.0286*

Following Lim et al. (2019) we attempted a single LSTM jointly predicting direction and magnitude via a turnover-penalised Sharpe-ratio loss, in six variants. All converge to the same qualitative failure - near-zero positions, mostly negative Sharpe. A sixth, warm-started from a cross-entropy direction pretrain, escapes the degenerate regime (mean $|\text{pos}|$ rises to 0.156) and produces 5/15 positive test Sharpes, yet still does not beat fixed momentum (mean ASR -0.287). The negative result survives, and its mechanism is now identified rather than guessed at: the flat failure mode was the initialisation, and fixing it is not sufficient at this data scale (Appendix S, Figure 17(c)).

Both learned baselines train successfully - this is not a convergence failure - and both underperform fixed momentum on mean ASR (-0.534 and -0.253 versus -0.007). They do beat it more often *per instrument* (4/15 and 6/15) than the mean comparison alone suggests, so the gap is driven by a few large losses rather than uniform underperformance. The magnitude rule is the same volatility-scaled baseline throughout; only the direction signal changes, so this is a controlled comparison in the same sense as the rest of the paper.

Multi-seed study. The learned baselines use a fixed seed (42) in the headline table. A 5-seed study (Table 21, right-hand columns) gives mean ASR -0.364 for logistic regression (average within-instrument standard deviation across seeds 0.581) and -0.105 for the MLP (std 0.426), versus fixed momentum's -0.007. The learned baselines stay negative on average, and the large within-seed

Table 19: Learned direction baselines versus fixed momentum, per instrument, single split (Appendix R). Magnitude is the same volatility-scaled baseline throughout; only the direction signal changes.

Instrument	Fixed momentum		Logistic		MLP	
	ASR	MDD	ASR	MDD	ASR	MDD
SPY	0.220	25.2	−1.061	48.5	0.777	20.2
QQQ	0.235	21.9	−0.726	38.9	0.736	22.8
IWM	−0.680	40.5	−0.259	24.0	0.138	27.3
GLD	0.505	19.4	−0.479	35.1	0.433	28.4
USO	−0.584	39.0	−0.458	33.6	−0.209	32.1
SLV	−1.019	50.2	0.912	20.0	0.432	20.1
TLT	−0.448	36.9	−1.650	61.0	−0.729	39.6
IEF	−0.107	28.6	−0.552	36.5	−0.486	30.8
BTC	0.062	37.6	−0.393	55.7	0.634	25.1
ETH	−0.318	31.8	−0.788	43.5	−0.448	31.5
SOL	1.105	26.9	−1.967	47.1	−2.261	51.3
EURUSD	0.180	32.2	−0.259	29.6	−0.113	28.8
JPYUSD	0.964	14.4	0.234	21.7	−1.706	62.7
GBPUSD	0.549	15.4	0.039	30.3	−0.145	34.7
AUDUSD	−0.766	47.0	−0.594	40.7	−0.847	51.8
mean	*−0.007*	*31.1*	*−0.534*	*37.8*	*−0.253*	*33.8*

standard deviation confirms the seed sensitivity we disclosed rather than dismissing it. T_1's computation, the gating mechanism and the direction signal are all fully deterministic given fixed data, so no such study is needed for the paper's headline result.

S DMN REPRODUCTION: FULL DETAIL

Following Lim et al. (2019) we attempted a single LSTM jointly predicting direction and magnitude via a turnover-penalised Sharpe-ratio loss. Six variants were run, each targeting a specific identified failure mode of the previous one rather than continuing an open-ended search.

Table 20: DMN reproduction attempts. Variants 1, 4, 5 and 6 were run on the full 15-instrument universe; variants 2 and 3 were run on a reduced 8-instrument subset and are therefore not directly comparable to the others. Per-instrument values for variants 1, 4 and 5 are in Table 21.

Variant	Change from the previous variant	Outcome
1. Per-instrument, raw returns	-	mean|pos| 0.045-0.233; 11/15 negative Sharpe
2. Shared cross-sectional, raw returns (8-instr. subset)	pool windows across instruments, matching the original's training protocol	*worse*: mean|pos| 0.018-0.034
3. Shared, vol-normalised returns (8-instr. subset)	normalise each instrument to comparable scale before pooling (raw-scale pooling was the suspected cause)	no change: 0.017-0.034
4. Shared, vol-normalised, no train-time turnover cost	remove the other suspected cause: a daily turnover penalty with no hold period trivially rewards staying flat	no change: mean|pos| 0.021-0.036; 13/15 negative Sharpe
5. Warm-started from a BCE direction pretrain, then Sharpe fine-tuned	attack the *initialisation*, the remaining candidate cause	escapes the degenerate regime: mean|pos| rises to 0.156, 5/15 positive test Sharpe (e.g. SPY $+0.826$, BTC $+0.898$), mean ASR -0.287
6. Warm-start plus an L_2 anchor to the frozen warm-start policy	prevent Sharpe fine-tuning from drifting back to the flat solution	same qualitative picture as variant 5

All variants use the same LSTM architecture (single layer, 32 hidden units, tanh-squashed linear head) and Adam optimisation. Variants 1-4 converge on the same qualitative failure - near-zero

Table 21: Deep Momentum Network reproduction and the multi-seed learned-direction study, per instrument (Appendix S). $|p|$ is the mean absolute position: values far below 1 indicate the degenerate near-flat solution. The warm-started variant escapes that regime and still does not beat fixed momentum's mean ASR of -0.007. The 5-seed columns report the within-instrument mean and standard deviation across seeds.

Instr.	DMN v1 (per-instr.)		DMN v4 (shared)		DMN v5 (warm-start)		Logistic, 5 seeds		MLP, 5 seeds	
	$\|p\|$	ASR	$\|p\|$	ASR	$\|p\|$	ASR	mean	std	mean	std
SPY	0.154	0.244	0.024	−0.168	n/a	0.840	−0.523	0.330	0.508	0.377
QQQ	0.110	−0.047	0.026	−0.441	n/a	−0.711	−0.676	0.398	0.320	0.580
IWM	0.147	0.032	0.026	−1.119	n/a	−0.031	−0.395	0.343	−0.171	0.492
GLD	0.092	0.244	0.028	−0.988	n/a	−0.922	−0.545	0.277	0.435	0.275
USO	0.140	−1.160	0.020	−0.777	n/a	−0.280	−0.648	0.403	−0.389	0.184
SLV	0.039	0.125	0.032	0.007	n/a	−0.006	−0.150	0.514	0.427	0.453
TLT	0.094	−0.996	0.032	−1.284	n/a	−0.491	−0.159	0.809	−0.612	0.594
IEF	0.178	−0.990	0.038	−1.052	n/a	−1.183	−0.414	0.494	−0.430	0.360
BTC	0.044	0.381	0.023	−0.820	n/a	0.898	−0.029	0.568	0.800	0.101
ETH	0.061	−0.133	0.027	−0.491	n/a	0.041	−0.290	0.475	0.350	0.365
SOL	0.105	−1.205	0.022	−1.715	n/a	−1.661	−0.589	1.694	−1.103	1.368
EURUSD	0.073	−0.241	0.026	−0.931	n/a	0.148	−0.218	0.677	−0.045	0.286
JPYUSD	0.115	0.462	0.027	−0.533	n/a	−0.768	−0.308	0.730	−0.874	0.145
GBPUSD	0.089	−0.212	0.026	−1.186	n/a	−0.133	0.017	0.344	−0.264	0.364
AUDUSD	0.085	−0.903	0.026	−1.744	n/a	−0.055	−0.526	0.663	−0.527	0.440
mean	*0.102*	*−0.293*	*0.027*	*−0.883*	*n/a*	*−0.288*	*−0.364*	*0.581*	*−0.105*	*0.426*

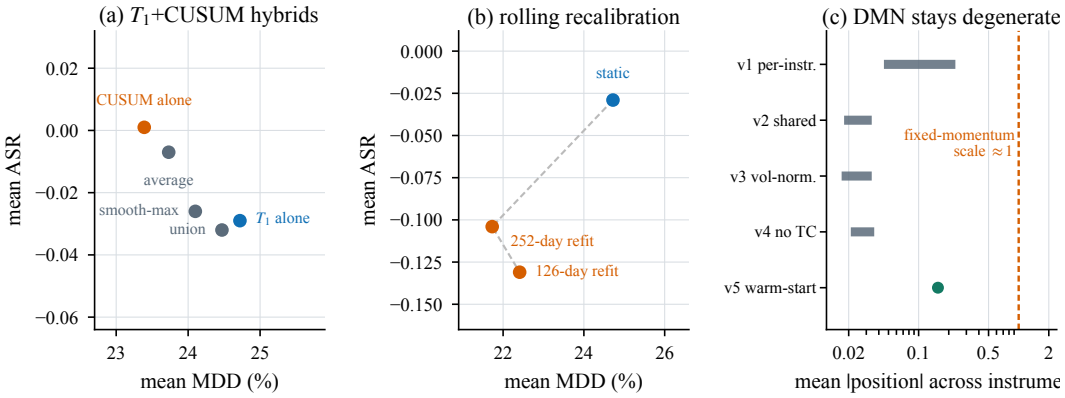


Figure 17: (a) Hybrid controllers combining T_1 and CUSUM: all three fixed rules land between their parents rather than beyond them. (b) Rolling threshold recalibration deepens the drawdown benefit and deepens its Sharpe cost with it. (c) The DMN reproduction across its variants: four converge to a near-flat position scale two orders of magnitude below the fixed-momentum reference, and warm-starting is what escapes it.

positions, mostly negative Sharpe - which rules out the two most plausible specific causes (raw-scale pooling, train-time turnover cost) as the dominant explanation: four mechanistically distinct fixes landing in the same band is itself the finding, not a reason to keep guessing. Direct Sharpe-ratio optimisation is a recognised hard case for gradient descent - noisy ratio-of-noisy-statistics gradients, and a randomly-initialised network's near-zero starting point is a genuine local optimum when the gradient signal toward larger, better-timed positions is weak.

Variant 5 confirms that diagnosis directly rather than leaving it a hypothesis: warm-starting from a cross-entropy direction pretrain escapes the flat regime entirely (mean$|$pos$|$ $0.021 \rightarrow 0.156$) and produces genuinely positive Sharpe ratios on a third of instruments - and *still* does not beat fixed momentum on average. The negative result survives with its mechanism now identified: the flat failure mode was the initialisation, and fixing the initialisation is not sufficient to make a Sharpe-loss-trained LSTM beat fixed momentum at this data scale. The original paper's result required decades of data across 88 futures contracts and substantial engineering that a reproduction at this scale should not be expected to match. We therefore report no learned-DMN row in Table 7; the classical volatility-scaling baseline we do use is DMN's own reported baseline, which keeps the comparison literature-grounded.

Table 22: Post-hoc extensions, aggregated over 15 instruments, single split (Appendix T). Baseline mean MDD 31.13%, mean ASR -0.007 throughout. No hybrid rule dominates either of its parents, and both recalibration schemes deepen the drawdown benefit at a larger Sharpe cost.

Configuration	ASR	MDD (%)	ΔMDD (pp)	improved
T_1 alone	-0.029	24.72	-6.42	13/15
CUSUM alone	0.001	23.39	-7.74	14/15
hybrid: union (max)	-0.032	24.47	-6.66	11/15
hybrid: average	-0.007	23.73	-7.40	13/15
hybrid: smooth-max (LSE)	-0.026	24.10	-7.03	12/15
Energy-T_1 + CUSUM:				
energy-T_1 alone	-0.142	22.85	-8.28	12/15
hybrid: union (max)	-0.088	24.46	-6.67	11/15
hybrid: average	-0.062	23.50	-7.63	11/15
hybrid: smooth-max (LSE)	-0.083	24.17	-6.96	11/15
Rolling recalibration:				
static threshold (default)	-0.029	24.72	-6.42	13/15
annual refit (252d)	-0.104	21.73	-9.41	12/15
semi-annual refit (126d)	-0.131	22.41	-8.72	12/15
Adaptive (ACI) threshold:				
static threshold (default), exceedance 0.331	-0.029	24.72	-6.42	13/15
ACI, lr $= 0.02$, exceedance 0.154	-0.080	29.80	-1.33	11/15
ACI, lr $= 0.05$, exceedance 0.122	-0.066	29.93	-1.20	11/15
ACI, lr $= 0.10$, exceedance 0.113	-0.066	29.73	-1.40	11/15

T FURTHER EXTENSIONS

Hybrid T_1/CUSUM controller. Both T_1 and CUSUM are non-dominated points in Table 7, so we tested a combined controller with three *pre-specified* fixed rules: union $z = \max(z_{T_1}, z_{\mathrm{CUSUM}})$, average $z = (z_{T_1} + z_{\mathrm{CUSUM}})/2$, and smooth-max (LogSumExp). Each combined score is z-scored and fed through the identical gate machinery. No rule dominates either single signal - the two scores are not strongly complementary on this evidence, and the average rule is a genuine compromise, not a win. Two structural facts about these rules are machine-checked (Appendix A): the union rule is exactly the pointwise *minimum* of the two individual gates, so it is the most conservative combination available; and LogSumExp is within $\log 2$ of the maximum, which is what justifies calling it a smooth max.

Energy-distance T_1 + CUSUM. The two non-dominated *competitor* signals (energy-T_1, best mean MDD; CUSUM, best mean ASR) are also not complementary: the same three hybrid rules land between the parents and are dominated by CUSUM on improved count and by energy-T_1 on mean MDD. Further support for the mechanism-class framing over any single construction.

Rolling threshold recalibration. The gate's $(\mu, \sigma, z_{\mathrm{thresh}})$ are re-estimated every 126 or 252 trading days on the expanding window, self-excluded and causally timed. Annual refits: mean MDD $24.72\% \to 21.73\%$ (mean ΔMDD -9.41 pp, 12/15 improved) at mean ASR -0.104; semi-annual refits reach MDD 22.41% at ASR -0.131. The insurance premium deepens when the threshold tracks the recent regime - a useful characterisation of the static-calibration limitation, not a fix that removes it.

Adaptive conformal-style threshold. A causally adaptive ACI-style threshold (Gibbs & Candès, 2021), $\theta_{t+1} = \theta_t + \mathrm{lr}\,(\mathbb{1}[z_t > \theta_t] - 0.10)$, restores nominal coverage: mean test-period exceedance falls from 33.1% (static) to $15.4\%/12.2\%/11.3\%$ for $\mathrm{lr} = 0.02/0.05/0.10$. That it does so is not luck - Theorem 2 shows the empirical exceedance rate is pinned to nominal at rate $O(1/T)$ for *any* score sequence, with no exchangeability assumption. But it erodes the drawdown benefit almost entirely: mean ΔMDD $-6.42 \to -1.33/-1.20/-1.40$ pp, MDD-improved $13/15 \to 11/15$, at mean ASR -0.066. The static threshold's $3\times$ exceedance is not a calibration bug - it *is* the mechanism, and forcing nominal coverage throws away the shared regime-shift signal the gate exists to catch. Coverage and risk-management benefit are in genuine tension on this data.

What remains open. Each of the three gaps this mechanism most obviously leaves - a hybrid controller, a portfolio-aware gate, and a distribution-shift-robust calibration - is tested directly above

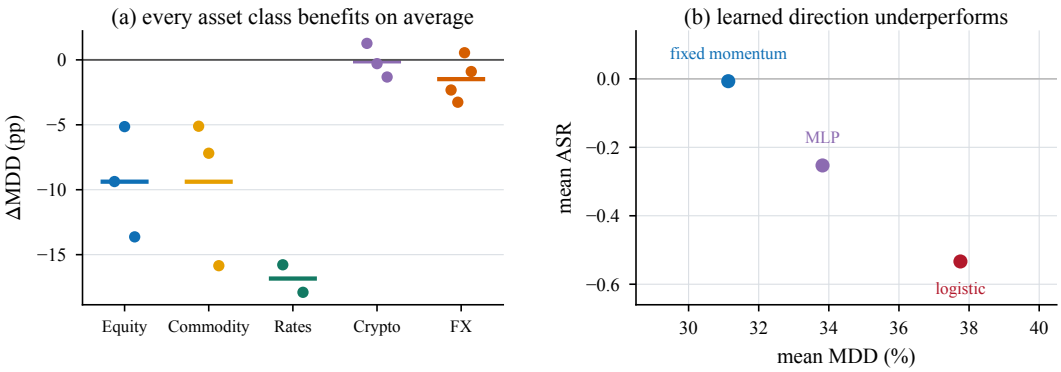


Figure 18: (a) ΔMDD by asset class; the horizontal bar is the class mean. Every class benefits on average, but the effect is concentrated in rates, equities and commodities, and is close to zero in crypto and FX. With two to three instruments per class this is suggestive, not a class-level generalisation claim (Section 5). (b) The learned-direction baselines against fixed momentum on both axes at once.

rather than left as an untested limitation. What remains genuinely open is a mechanism that moves further along the decorrelation/benefit frontier of Figure 15(b), and a theoretical account of when calibration coverage and drawdown control can be obtained together at all.

U Asset-class breakdown

V Hyperparameters

Table 23: Every hyperparameter used in the paper's headline configuration. None was tuned on test-period data; the sweeps of Appendix J report sensitivity around these values rather than selecting among them.

Quantity	Value	Quantity	Value
Return window w	20 days	Volatility target σ_{target}	15% ann.
Neighbours k	20	Volatility window	60 days
Softmax temperature τ	0.5	Max leverage $m_{\max}$	2.0
Gate scale s	2.5	One-way transaction cost	10 bps
Gate floor $g_{\min}$	0.2	MACD kernel pairs	(8,24), (16,48), (32,96)
Calibration percentile	90th	Crisis rolling-high lookback	252 days
Train/test split	60/40 chron.	Crisis threshold percentile	80th (train)
Walk-forward folds	5, expanding	Learned-direction label horizon	5 days
Backtest warm-up	252 days	Minimum train days (portfolio)	300
T_1 calibration start index	200	Evaluation window	2014-01-01 to 2024-12-31

W Reproducibility details

Instrument universe and data source. All 15 instruments (SPY, QQQ, IWM, GLD, USO, SLV, TLT, IEF, BTC-USD, ETH-USD, SOL-USD, EURUSD=X, JPY=X, GBPUSD=X, AUDUSD=X) are adjusted-close daily OHLCV series obtained from Yahoo Finance via the `yfinance` package over 2014-01-01 to 2024-12-31, except where a longer window is stated explicitly (Section 4.4). SOL-USD's history begins 2020-04-10; all others cover the full window. No survivorship-bias concern applies to single-ticker ETF, FX and crypto instruments in the way it would to an equity index constituent list. Every hyperparameter needed to reproduce the pipeline is in Appendix V, the two algorithms are given in full in the supplementary material, and each experiment's design is stated with its result.

How the numbers in this paper were produced. Every experiment writes a raw result file, and every table and figure in this paper is generated from those files by script: no number is hand-transcribed, and a checking script re-derives each of the 76 numeric claims in the text from the raw outputs and fails if any one of them does not appear verbatim. We mention this because it is the

mechanism that keeps a paper of this size consistent with its own data, not because the scripts are novel.

Compute. Every experiment except the DMN variants runs on CPU, and the full suite completes in well under an hour of wall-clock time on a standard cloud CPU instance. The DMN reproduction uses PyTorch; each variant trains in under 90 seconds, and a GPU is convenient but not required.

Randomness. T_1's computation, the gating mechanism and both direction signals are fully deterministic given fixed data - there is no random-seed dependence in this paper's headline result. The learned baselines use a fixed seed (42), with a 5-seed study in Appendix R.1; the block bootstrap of Section 4.3 uses a fixed seed (0) and 20,000 draws.

Formal verification. The Lean 4 development that machine-checks every mathematical statement in this paper is included with this submission as `PaperMath.lean`, together with the pinned toolchain and Mathlib revision. Running `lake exe cache get && lake env lean PaperMath.lean` re-checks all 23 results and emits the axiom audit reproduced in Appendix A. It requires no market data and no part of the experimental pipeline.

Code and data. The experiment code and cached instrument data are not included with this submission. They will be released publicly, with the exact commands that regenerate each table and figure, on publication.