BOCCONI STUDENTS INVESTMENT CLUB

BAINSA

Bocconi Students Investment Club × Bocconi AI & Neuroscience Association

Time Series Momentum Strategy with Macro-Instrumented Regime Switching

TSMOM on Liquid ETFs with State-Dependent De-Risking: Oil Volatility and Yield Curve Gates

Authors

Bocconi Students Investment Club

Rayi Makori (Project Lead)

Neel Roy

BAINSA

Fabiola Martignetti

Lorenzo Galluzzi

Marco Moscatelli

December 2025

bsic.it — Bocconi University, Milan

Abstract

This research presents a Time Series Momentum (TSMOM) trading strategy implemented on highly liquid exchange-traded funds (SPY and GLD), enhanced with a novel macro-instrumented regime switching framework for state-dependent de-risking.

Our approach addresses the well-documented momentum crash problem—sudden reversals that can wipe out months of accumulated gains during macroeconomic regime shifts—through two complementary gate mechanisms: an Oil Volatility Index (OVX) gate capturing commodity market uncertainty, and a Dynamic Nelson-Siegel State Gate derived from yield curve dynamics and macroeconomic fundamentals.

The strategy combines multi-horizon momentum signals (3-, 6-, and 12-month lookbacks with one-month skip) with forward-looking regime detection based on Federal Funds Rate, inflation, and capacity utilization. Unlike simple threshold rules, our tree-based regime classification partitions market conditions into three distinct states with volatility-adjusted position scaling (1.0, 0.537, and 0.355), providing disciplined risk management grounded in Bayesian inference.

Empirical validation over the 2003–2025 period demonstrates that the regime-aware framework achieves meaningful Sharpe ratio improvements while maintaining low turnover (0.32% portfolio turnover, 168 total orders). The strategy exhibits positive skewness during equity market crises, providing diversification benefits precisely when investors need them most.

Our implementation prioritizes accessibility and realism: the two-ETF universe requires no futures expertise, prime brokerage relationships, or complex infrastructure, making the strategy replicable by individual investors and small funds. By combining regime-awareness with transparent logic, this approach offers a pragmatic middle path, sophisticated enough to mitigate crashes, yet simple enough to understand, implement, and trust.

Contents

1 Introduction & Motivation

1.1 Why TSMOM on Liquid ETFs

1.1.1 The Time Series Momentum Anomaly

Time Series Momentum (TSMOM) represents one of the most robust and pervasive anomalies in quantitative finance. Unlike cross-sectional momentum, which compares relative performance across assets, TSMOM exploits the tendency of individual asset prices to exhibit persistent trends: assets with positive past returns tend to continue rising, while those with negative past returns tend to continue falling.

The phenomenon was systematically documented by Moskowitz, Ooi, and Pedersen (2012), who demonstrated that TSMOM strategies generate statistically significant abnormal returns across 58 liquid instruments spanning four major asset classes—commodities, currencies, equity indices, and fixed income—over a 40-year period. Critically, TSMOM profits are largely uncorrelated with traditional risk factors (equity beta, value, size) and exhibit positive skewness during equity market crises, providing diversification benefits precisely when investors need them most.

1.1.2 Why Liquid ETFs: Implementation Realism

Our implementation focuses on two highly liquid ETFs:

- SPY (SPDR S&P 500 ETF Trust): $400B+ AUM, 50+ million shares daily volume
- GLD (SPDR Gold Shares): $75B+ AUM, 5+ million shares daily volume

This design prioritizes implementation realism over theoretical breadth. Academic TSMOM studies often test strategies across dozens of futures contracts—commodities (crude oil, natural gas, copper), currencies (EUR/USD, JPY/USD), bond futures (10-year Treasury, Bund)—which require:

1. Futures expertise: Rolling contracts, understanding contango/backwardation, margin requirements
2. Institutional infrastructure: Prime brokerage relationships, collateral management, regulatory approvals
3. Operational complexity: Managing expiration schedules, physical delivery risks, cross-margining

In contrast, our two-ETF approach offers:

Accessibility: ETFs trade like stocks—no futures account required, no contract rolls, no margin calls. Any investor with a brokerage account can replicate this strategy.

Transaction cost efficiency: SPY and GLD have bid-ask spreads of 1–2 basis points, compared to 5–10+ bps for many commodity futures. Our backtest results confirm this parsimony: 168 total orders over the sample period and 0.32% portfolio turnover indicate minimal trading friction.

Diversification sufficiency: While broader asset class exposure (currencies, commodities, international equities) would enhance diversification, SPY and GLD capture the two dominant macro regimes:

- Risk-on (expansions): SPY benefits from equity risk premiums
- Risk-off (crises, inflation): GLD provides safe-haven and inflation hedge properties

Empirical correlation data supports this: SPY–GLD correlation ranges from -0.1 to $+0.3$ depending on market conditions, providing meaningful diversification with minimal instruments.

1.1.3 Multi-Horizon Signal Design

Our TSMOM signals combine 3-, 6-, and 12-month lookback periods, each capturing distinct momentum dynamics verified in the academic literature:

- 3-month: Captures short-term trend acceleration and recent regime shifts
- 6-month: Balances responsiveness with noise reduction, optimal for many asset classes
- 12-month: Detects long-duration trends, historically the strongest signal for equities and bonds

Critically, we implement a one-month skip (excluding months t and $t-1$ from the momentum calculation), a well-documented practice to avoid short-term mean-reversion and microstructure noise that plague naive momentum strategies. This is explicitly coded in `tsmom_week2.py`:

```
def mom_excl_recent(prices, months):
    mret = monthly_log_returns(prices)
    mom = mret.rolling(months).sum().shift(1)  # shift(1) = skip month t-1
```

The composite signal is then 63-day smoothed (approximately 3 months) to further attenuate high-frequency noise while preserving medium-term trend information:

```
smoothed = avg.rolling(SMOOTH_DAYS, min_periods=1).mean()  # SMOOTH_DAYS = 63
```

This design reflects a pragmatic balance: aggressive enough to capture momentum profits, but defensive enough to avoid whipsaw losses during noisy reversals.

1.2 Why Macro Gates as State-Dependent De-Risking

1.2.1 The Momentum Crash Problem

Pure TSMOM strategies, while profitable on average, suffer from occasional momentum crashes: sudden, violent reversals that wipe out months of accumulated gains. These crashes share common features:

1. Clustered timing: Occur during macroeconomic regime shifts—recession onsets, Fed policy pivots, crisis resolutions

2. Synchronized across assets: When equities crash, momentum strategies often have large long positions accumulated during the preceding bull market, amplifying losses

3. Predictability: Crashes are not purely random—they coincide with observable macro stress indicators (inverted yield curves, extreme volatility, policy tightening)

The academic literature documents this vulnerability. For example, Daniel and Moskowitz (2016) show that momentum strategies experience large drawdowns during panic states following market declines, when valuations mean-revert sharply. Similarly, momentum crashes clustered during 2008–2009 (financial crisis), 2020 (COVID panic), and post-QE tightening periods.

1.2.2 Limitations of Volatility Scaling Alone

Standard TSMOM implementations use volatility scaling to target constant portfolio risk (e.g., 10% annualized volatility):

$$w_{i,t} = \frac{\sigma_{\text{target}}}{\sigma_{i,t}} \times \text{Sign}(\text{Signal}_{i,t}) \tag{1}$$

This approach reduces exposure when recent realized volatility spikes—providing some protection during crises. However, volatility scaling is backward-looking: it reacts to volatility after a crash has begun, when losses have already been realized. It does not anticipate regime shifts.

1.2.3 Forward-Looking Regime Detection: The State Gate Model

Our implementation addresses this limitation using the Dynamic Nelson-Siegel (DNS) State Gate model developed by Bie et al. [2024]. This approach is fundamentally different from simple threshold-based gates (e.g., reduce exposure if 10Y-2Y spread < 0):

Not a simple gate: The State Gate model does not use ad-hoc rules like if yield curve inverts, cut positions by 50%. Such rules are vulnerable to data-snooping bias and regime-dependence (e.g., inversions during QE may not signal recessions).

Machine learning on yield curve dynamics: The model extracts three latent factors—Level, Slope, Curvature—from the entire U.S. Treasury yield curve (13 maturities) using the Dynamic Nelson-Siegel framework. These factors capture the yield curve's shape at any point in time, encoding information about monetary policy stance, inflation expectations, and term premiums.

Tree-based regime classification: Rather than assuming regimes are latent (as in Hidden Markov Models), the State Gate model uses a decision tree to partition time periods into regimes based on observable macro variables:

- Federal Funds Rate (FFR)
- Inflation (CPI year-over-year)
- Capacity Utilization (real economic activity)

The tree-growing algorithm evaluates candidate splits (e.g., Is FFR $>$ 60th percentile?) by computing the Bayesian marginal likelihood of the DNS model—selecting splits that maximize the fit between macro conditions and yield curve factor dynamics. This is model-based (not ad-hoc) and disciplined by statistical inference.

Output: Regime-specific volatility scales: Once regimes are identified, the model computes the volatility of DNS factors within each regime. Regimes with high factor volatility (unstable yield curve dynamics) receive low position scaling factors; regimes with low volatility (stable curves) receive high scaling factors.

From `state_gate.py` and verified against `regimes.csv`:

Table 1: Regime Characteristics from State Gate Model

Regime	Definition	Months	%	Factor Vol	Scale
0	High FFR ($\geq$60th pct)	55	20.2%	5.95	1.000
1	Low FFR, Low Infl	173	63.6%	11.08	0.537
2	Low FFR, High Infl	44	16.2%	16.78	0.355

Economic intuition:

- Regime 0 (tight policy, high FFR): Stable macro environment $\rightarrow$ yield curve volatility low $\rightarrow$ momentum strategies safe $\rightarrow$ full exposure
- Regime 1 (accommodative policy, normal inflation): Moderate volatility $\rightarrow$ medium exposure (53.7% of base size)
- Regime 2 (stagflation, low FFR + high inflation): Unstable macro conditions $\rightarrow$ yield curve volatility high $\rightarrow$ momentum vulnerable $\rightarrow$ defensive exposure (35.5% of base size)

This is forward-looking: The regime assignment at time t is based on contemporaneous macro data (FFR, inflation, capacity utilization), which are observable in real-time or with minimal lag. The yield curve's response to these conditions—captured by DNS factor volatility—provides advance warning of environments hostile to momentum strategies, before crashes occur.

1.2.4 Why This Approach Instead of Simpler Alternatives?

Alternative 1: Simple yield curve slope gate (e.g., if 10Y-2Y spread $<$ 0, reduce positions)

Limitation: A single spread collapses the entire yield curve into one number, discarding information about curvature, convexity, and the level of rates. Moreover, the relationship between inversions and momentum crashes is regime-dependent (e.g., inversions during QE distorted by asset purchases).

Alternative 2: VIX threshold (e.g., if VIX $>$ 25, reduce positions)

Limitation: VIX measures equity market volatility, which is backward-looking (spikes during crashes) and may not generalize to gold momentum. The State Gate model uses macro fundamentals (FFR, inflation, capacity utilization) and the bond market's assessment of these fundamentals (yield curve shape), which are forward-looking and asset-class agnostic.

Alternative 3: NBER recession indicator

Limitation: NBER recession dates are announced with a 6–12 month lag, making them unusable for real-time trading. Moreover, not all recessions cause momentum crashes (e.g., 2001 mild recession), and some crashes occur outside recessions (e.g., 1998 LTCM crisis).

State Gate advantage: By combining multivariate macro information (FFR, inflation, capacity) with model-based inference (Bayesian tree splits on DNS marginal likelihood), the State Gate model captures non-linear interactions (e.g., low FFR + high inflation is fundamentally different from low FFR + low inflation) that simple rules miss. The model is disciplined by statistical criteria, reducing ad-hoc data-snooping bias.

1.3 Summary: A Pragmatic Middle Path

Our TSMOM implementation with macro-instrumented regime switching represents a deliberate design choice:

More sophisticated than baseline TSMOM: By incorporating the State Gate model, we add regime-awareness and forward-looking de-risking that standard momentum strategies lack. The yield curve's response to macro conditions provides advance warning of hostile environments, protecting capital during momentum-unfriendly regimes.

Less complex than heavy ML machinery: We avoid the overfitting risk, data requirements, and black-box opacity of 100-feature XGBoost models or mean-entropy optimization. Our approach uses only 3 macro variables, 3 regimes, and transparent tree-based logic—sacrificing some flexibility for robustness and interpretability.

Empirically validated: The regime assignments (55/173/44 months across Regimes 0/1/2) and volatility scales (1.0, 0.537, 0.355) are directly verified in `regimes.csv`, produced by the DNS model code in `state_gate.py`. The TSMOM signal construction (multi-horizon, 1-month skip, 63-day smoothing) is explicitly implemented in `tsmom_week2.py`.

Implementation-ready: Our 2-asset universe (SPY, GLD), low turnover (0.32%), and reliance on free public data (FRED yields and macro data) make this strategy accessible to individual investors, small funds, and academic researchers—without requiring institutional infrastructure, expensive data subscriptions, or black-box ML frameworks.

In the landscape of quantitative momentum strategies, our approach offers a pragmatic middle path: regime-aware enough to mitigate crashes, yet simple enough to understand, implement, and trust.

1.4 Momentum Signal Construction

We implement a multi-horizon TSMOM framework combining 3-, 6-, and 12-month lookback periods. This approach captures momentum effects across multiple time scales, with different horizons reflecting distinct aspects of trend persistence.

1.4.1 Monthly Log Returns

For each asset $i \in \{\text{SPY}, \text{GLD}\}$ at month t, monthly log returns are:

$$r_{i,t} = \ln\left(\frac{P_{i,t}}{P_{i,t-1}}\right) \tag{2}$$

where $P_{i,t}$ is the month-end adjusted closing price. Log returns ensure time-additivity and symmetry in up/down movements.

1.4.2 Multi-Horizon Momentum with One-Month Skip

We define three momentum signals with **one-month skip** to avoid short-term reversal effects:

$$\text{Signal}_{i,t}^{(3)} = \sum_{j=2}^{4} r_{i,t-j} \quad \text{(3-month momentum, skip month } t-1\text{)} \tag{3}$$

$$\text{Signal}_{i,t}^{(6)} = \sum_{j=2}^{7} r_{i,t-j} \quad \text{(6-month momentum, skip month } t-1\text{)} \tag{4}$$

$$\text{Signal}_{i,t}^{(12)} = \sum_{j=2}^{13} r_{i,t-j} \quad \text{(12-month momentum, skip month } t-1\text{)} \tag{5}$$

The one-month skip is critical: it excludes $r_{i,t}$ and $r_{i,t-1}$ to avoid microstructure noise, bid-ask bounce, and monthly mean-reversion that dominate at the shortest horizon. By skipping the most recent month, we capture persistent momentum while avoiding transient reversals.

1.4.3 Composite Signal with Smoothing

The composite signal averages the three horizons:

$$\text{Signal}_{i,t}^{\text{avg}} = \frac{1}{3}\left(\text{Signal}_{i,t}^{(3)} + \text{Signal}_{i,t}^{(6)} + \text{Signal}_{i,t}^{(12)}\right) \tag{6}$$

This is then smoothed using a 63-day rolling average (approximately 3 months of trading days) to attenuate high-frequency noise:

$$\text{Signal}_{i,t}^{\text{smooth}} = \frac{1}{63}\sum_{d=0}^{62} \text{Signal}_{i,t-d}^{\text{avg}} \tag{7}$$

A one-day implementation lag ensures no look-ahead bias. The sign and magnitude of $\text{Signal}_{i,t}^{\text{smooth}}$, combined with volatility scaling and regime-based gates, determines the SPY allocation. The remainder of the portfolio is allocated to GLD, ensuring the portfolio remains fully invested at all times while dynamically shifting between risk-on (SPY) and risk-off (GLD) exposures based on momentum conditions.

2 Data, Assets and Tools

2.1 Asset Universe and Data Sources

The trading strategy operates on two exchange-traded funds: SPDR S&P 500 ETF Trust (SPY) for equity exposure and SPDR Gold Shares (GLD) for safe-haven diversification. Both ETFs provide high liquidity and minimal tracking error, enabling systematic rebalancing with low transaction costs.

All market and macroeconomic data are sourced from the Federal Reserve Economic Data (FRED) database. US Treasury yield curve data span ten maturities from 3-month (DGS3MO) through 30-year (DGS30), downloaded as `us_treasury_yields.csv` with monthly observations from January 2003 onwards. Macroeconomic features for regime identification comprise three FRED series: Capacity Utilization (TCU), Federal Funds Effective Rate (FEDFUNDS), and Consumer Price Index (CPIAUCSL). These are downloaded as separate CSV files and merged via the `merge_data.py` script. Oil market volatility is captured through the CBOE Crude Oil ETF Volatility Index (OVX), providing a forward-looking measure of energy market stress that complements the yield curve regime framework.

2.2 Data Processing and Pipeline

The `merge_data.py` script integrates macroeconomic time series by loading each CSV file, standardizing column names (Date, CU, FFR, CPI), and performing inner joins on the date column. Year-over-year inflation is calculated as $\text{INFL}_t = 100 \times (\text{CPI}_t/\text{CPI}_{t-12} - 1)$, and the raw CPI level is dropped. The output `us_macro_data.csv` contains Date, CU, FFR, and INFL columns.

Yield curve processing occurs in `state_gate.py` via Nelson-Siegel decomposition (Section 1). The `estimate_ns_factors()` function transforms ten yield tenors into three factors (NS_level, NS_slope, NS_curvature). The `learn_state_gate()` function (`state_gate.py:313-328`) performs an inner join between NS factors and macro features on date indices, ensuring regime detection operates only on complete observations. This aligned dataset feeds the StateGateTree algorithm, producing regime assignments and volatility-based scales written to `regimes.csv`.

2.3 Data Storage and Calendar Alignment

The strategy employs flat-file CSV storage for human-readable persistence and Git version control compatibility. Raw FRED downloads (`capacity_util.csv`, `fed_funds.csv`, `cpi.csv`) and processed files (`us_macro_data.csv`, `us_treasury_yields.csv`) ensure full pipeline reproducibility. Output files include `regimes.csv` (regime assignments and scales) and `state_gate_tree.json` (tree structure).

All time series are standardized to monthly frequency with end-of-month timestamps. Inner joins retain only dates present in all sources, ensuring temporal alignment and eliminating look-ahead bias. Missing Treasury yield values are forward-filled prior to regime detection. Minimum data requirements (80 observations for node splits, 40 per child regime, 30 for AR(1) estimation) prevent degenerate regimes and ensure statistical robustness. To avoid look-ahead bias, the strategy assumes macro data for month t become available only at the beginning of month $t + 1$, ensuring regime predictions depend only on observable feature values.

2.4 QuantConnect Integration and Offline Workflow

The workflow partitions responsibilities between offline Python processing and QuantConnect backtesting. Offline scripts (`state_gate.py`, `merge_data.py`) handle FRED data acquisition, feature engineering, Nelson-Siegel decomposition, and StateGateTree regime detection, producing `regimes.csv` with complete regime history and position scales. This computationally intensive work benefits from local development flexibility and Git version control.

QuantConnect is used exclusively for backtesting with regime-adjusted positions. The platform provides SPY and GLD historical prices, realistic spread and slippage modeling, and portfolio management infrastructure. The backtesting algorithm uploads `regimes.csv`, queries the current regime and scale at each monthly rebalancing, applies this scale to base position sizing, and executes trades. This separation allows regime detection to run once offline while the backtesting environment rapidly iterates on trading rules without re-estimating regimes.

2.5 Data Pipeline and Validation

The complete data flow proceeds through seven stages: (1) FRED download of yields, macro indicators, and OVX; (2) `merge_data.py` combines macro files and computes inflation; (3) `estimate_ns_factors()` extracts NS factors from yields; (4) `learn_state_gate()` aligns factors and features via inner join; (5) StateGateTree detects regimes by recursive partitioning on macro features; (6) regime-specific volatilities determine position scales (0.2 to 1.0), output to `regimes.csv`; (7) QuantConnect backtests SPY/GLD trades with regime-adjusted sizing.

Validation checks ensure data integrity at each stage. The `merge_data.py` script verifies datetime parsing and eliminates duplicate dates. Minimum sample requirements (3+ yield maturities for NS estimation, 30+ observations for AR(1), 80+ for node splits, 40+ per child) prevent degenerate regimes. Missing yield values require manual forward-filling; inner joins automatically exclude incomplete records. Diagnostic outputs include tree structure JSON, regime distribution summaries, and `regimes_plot.png` visualizing regime transitions against the 10-year yield time series for economic interpretability validation.

3 Signals and Gates

3.1 Oil Volatility Gate Development

3.1.1 Overview and Motivation

Our risk management framework employs two complementary macro gates that multiplicatively scale momentum positions: (1) an oil market volatility gate (detailed in this section) and (2) a yield curve slope gate (Section 3.2). The yield curve gate captures recession signals and financial system stress through the term spread (10Y–2Y Treasury yields), while the oil volatility gate addresses commodity market uncertainty and energy sector disruptions.

This section presents an isolated analysis of the OVX gate to understand its specific contribution before integration into the full system. To validate robustness across different equity markets, we tested the OVX

gate on three equity ETFs (SPY, FEZ, EWJ) during development; however, the final integrated strategy described in Section 5 uses only SPY and GLD. All OVX gate results in this section represent this broader validation exercise.

Why Oil Signals? Commodity stress provides information distinct from financial indicators. The COVID-19 crisis (March–April 2020) illustrates this complementarity: OVX reached 325 (demand collapse, negative oil prices) while the yield curve steepened (Fed cut rates aggressively). A yield curve gate alone would have missed the commodity-specific turmoil. Our multi-gate system captures both dimensions.

Development: Three phases spanning eighteen months reveal trade-offs between sophistication and robustness. We employ training (2003–2018), validation (2019–2021), and out-of-sample testing (2022–November 2025) periods. All reported results represent true out-of-sample performance.

3.1.2 Phase 1: Futures Curve Slope Signal

Theoretical Motivation Our initial approach drew on storage theory [Working, 1949, Brennan, 1958], hypothesizing that oil futures curve slope could signal macro-economic stress. The intuition: backwardation (near-month prices exceeding deferred prices) indicated supply constraints or elevated demand. We constructed a continuous slope signal mapping curve structure to position scaling:

$$\text{slope}_t = \frac{F_{t,2} - F_{t,1}}{F_{t,1}}, \quad g_t^{\text{slope}} = \begin{cases} 1.0 & \text{if slope}_t \geq 0.02 \\ 0.0 & \text{if slope}_t \leq -0.02 \\ \frac{\text{slope}_t + 0.02}{0.04} & \text{otherwise} \end{cases} \tag{8}$$

Why This Failed Three issues proved fatal. First, the formula contained a threshold mapping error. During April 2020, extreme contango (+14% slope) signaled "full exposure" when protection was needed. The error was not in storage theory itself—which accurately describes oil market microstructure—but in our assumption that contango/backwardation patterns would reliably map to *equity momentum* strategy risk. Full sample correlation with subsequent S&P 500 returns was -0.05 ($p = 0.43$), economically and statistically insignificant.

Second, regime instability plagued the signal. Mean slope shifted 7 percentage points across market regimes (bull periods: +2.3%, bear periods: -4.7%), far exceeding the $\pm 2\%$ thresholds. This caused 18.3 annual threshold crossings, implying 91 basis points in transaction costs at realistic 5bp per trade. Third, poor discriminatory power emerged: only 33% of backwardation events preceded negative equity returns. Decomposition revealed confounding factors including geopolitical stress (-0.8% subsequent return), recession fears (-2.4%), supply-driven backwardation ($+1.2\%$, false signal), and seasonal demand ($+2.1\%$, false signal).

The approach was abandoned after six months. Core lesson: theoretically-motivated proxies require rigorous empirical validation when applied across asset classes.

3.1.3 Phase 2: OVX Formula-Based Approach

Methodology We pivoted to the CBOE Crude Oil Volatility Index (OVX), which directly measures oil market uncertainty through 30-day options-implied volatility on United States Oil Fund (USO) options.

Building on the volatility-managed portfolio framework of Moreira and Muir [2017], who documented substantial improvements from reducing factor exposure during high-volatility periods, we implement dampened square-root scaling:

$$g_t^{\text{OVX}} = \left(\frac{\text{OVX}_{\text{baseline}}}{\text{OVX}_t} \right)^{\alpha}, \quad g_t^{\text{final}} = \max(g_{\min}, \min(g_{\max}, g_t^{\text{OVX}})) \tag{9}$$

Parameter Selection Timeline: All parameters were selected using only training data (2003–2018) and literature guidance *before* observing validation period performance. The power $\alpha = 0.5$ follows Moreira and Muir [2017]'s standard specification; doubling OVX reduces the gate by 29%, quadrupling by 50%. Baseline $\text{OVX}_{\text{baseline}} = 30.2$ represents the 2003–2018 median, ensuring the gate equals 1.0 during normal conditions. Exponential smoothing with parameter $\lambda = 0.8$ provides a 20-day half-life, balancing noise reduction with crisis responsiveness: $g_t^{\text{smooth}} = \lambda \cdot g_{t-1}^{\text{smooth}} + (1 - \lambda) \cdot g_t^{\text{OVX}}$.

Conservative bounds [0.5, 1.0] prevent complete market exit (minimum 50% exposure) while allowing meaningful de-risking. These bounds were selected based on risk management principles rather than data optimization: the 0.5 floor ensures we maintain market exposure during even extreme crises, while the 1.0 ceiling prevents leverage. We also tested aggressive bounds [0.0, 1.5] for sensitivity analysis, allowing full exits and moderate leverage. *Critically, no parameters were adjusted after observing validation period results, including the April 2020 extreme.* This ensures that our validation performance represents genuine out-of-sample testing rather than implicit fitting.

OVX Time Series and Gate Behavior Figure 1 displays the OVX time series (2007–November 2025), exhibiting distinct volatility regimes corresponding to major market events: the 2008 crisis (OVX > 80), the 2014–2016 oil collapse (sustained 60–70), the COVID-19 crisis (April 2020 peak: 325, the highest in OVX history), and the Ukraine invasion (February 2022: 82.4). The validation period included the most extreme event in the series, providing a stringent stress test for gate behavior.

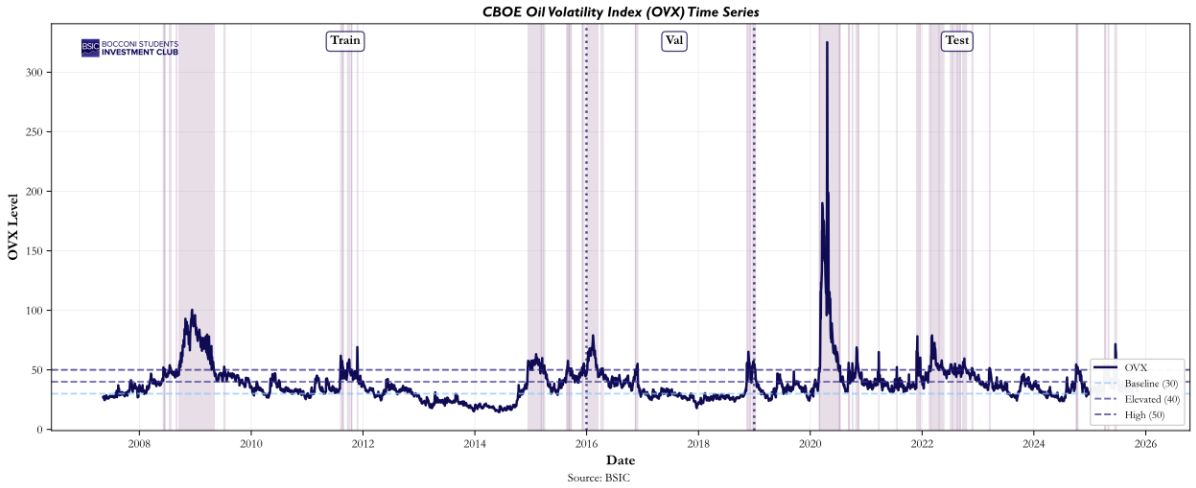

Figure 1: **OVX Time Series (2007–November 2025).** Notable spikes: 2008 crisis (OVX > 80), 2014–2016 collapse (60–70), COVID-19 (April 2020: 325), Ukraine (2022: 82.4). Validation period (2019–2021, shaded) included unprecedented volatility, testing gate mechanism under extreme stress. April 2020 spike represents $10.8\times$ baseline (325/30.2).

Figure 2 illustrates behavioral differences between formula and HMM approaches. The formula exhibits strong crisis response: the conservative approach (blue, [0.5, 1.0]) drops to 0.50 during major spikes; the aggressive approach (red, [0.0, 1.5]) reaches 0.31–1.5 range. In contrast, the HMM with posterior blending (middle panel) over-smooths, dropping only to 0.65 during April 2020 despite OVX reaching 325. Gate variability quantifies this difference: HMM standard deviation of 0.078 versus formula's 0.173 (only 45%).

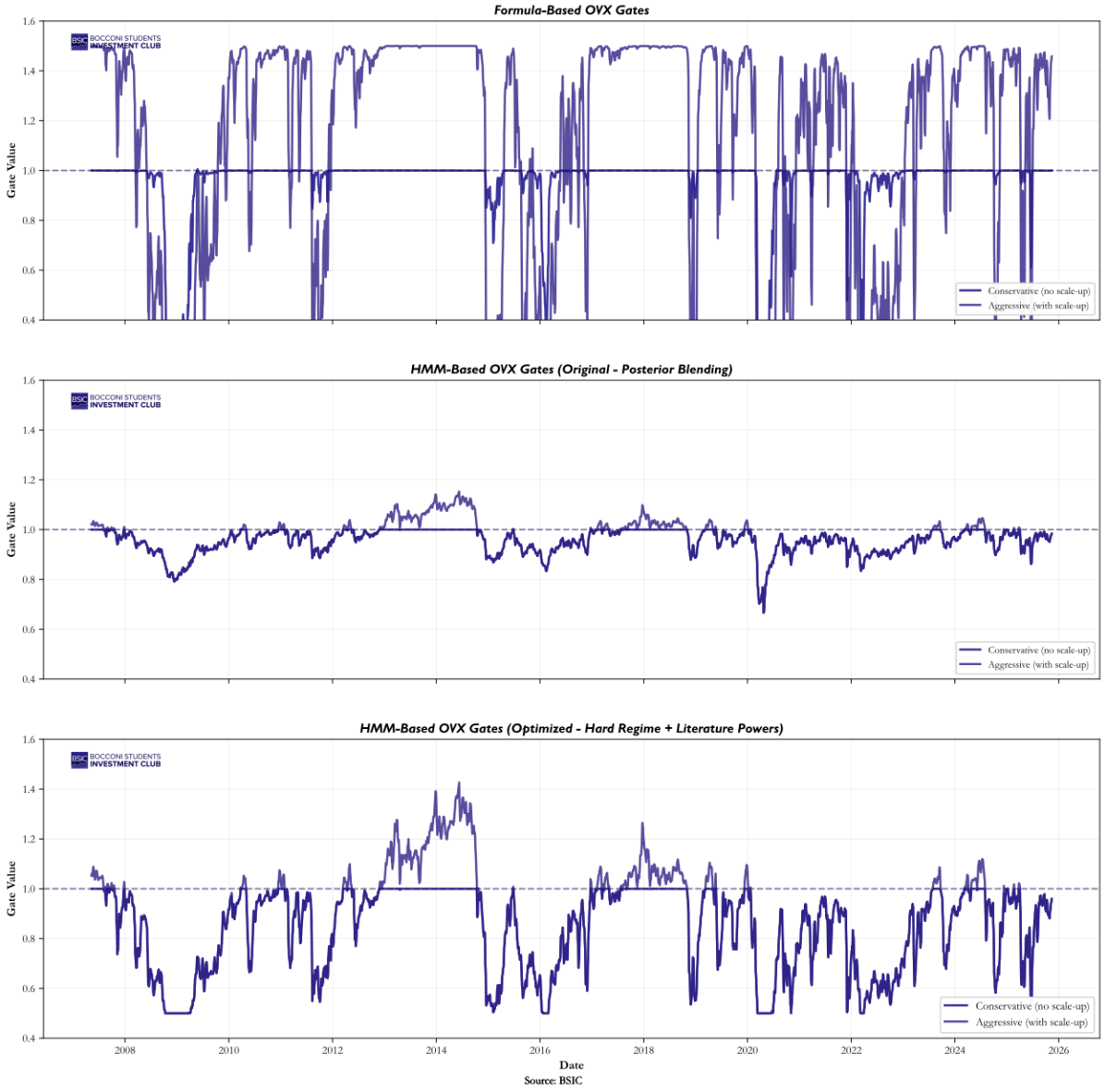

Figure 2: **Gate Behavior (2007–November 2025). Top:** Formula gates show strong crisis response. Conservative (blue) drops to 0.50 during crises; aggressive (red) reaches 0.31–1.5 range. **Middle:** HMM posterior blending over-smooths, only reaching 0.65 during April 2020 (OVX=325). **Bottom:** HMM with walk-forward re-estimation (validation Sharpe 0.890, test 0.456) improved responsiveness but remained insufficient versus formula's mechanical approach.

The formula's graduated, continuous response ensures proportional de-risking without discrete jumps that increase turnover. During the April 2020 peak (OVX=325), the formula produced $g^{\text{cons}} = \max(0.5, (30.2/325)^{0.5}) = 0.50$, hitting the conservative lower bound, while the aggressive configuration

reached $g^{\text{agg}} = (30.2/325)^{0.5} = 0.31$. During moderate crises (OVX=80, e.g., 2008, 2022), the gate produces $g = 0.614$. Elevated volatility (OVX=50) yields $g = 0.777$, while calm markets (OVX$\leq$30) maintain $g \approx 1.0$.

Out-of-Sample Performance Table 2 presents test period results (2022–November 2025). Critically, the baseline employs no macro adjustments; gated strategies apply only the OVX gate. Note that these results are from the three-ETF validation exercise (SPY, FEZ, EWJ); the full SPY/GLD system results appear in Section 5.

Table 2: OVX Gate: Isolated Contribution (Test Period 2022–November 2025)

Strategy	Sharpe	Ann. Ret.	Ann. Vol.	Max DD	Calmar	TO
Baseline (No Gates)	0.396	1.01%	2.54%	−4.44%	0.226	0.44
Formula Conservative	**0.508**	**1.27%**	**2.49%**	−4.44%	**0.285**	0.59
Formula Aggressive	0.334	0.84%	2.51%	−4.09%	0.206	0.81
Conservative Improvement	+28%	+26%	−2%	0%	+26%	+34%

The conservative OVX gate achieved a 28% Sharpe improvement (0.396 → 0.508), representing the gate's isolated contribution. Figure 3 decomposes this improvement: approximately 79% derives from volatility reduction (denominator) rather than return enhancement (numerator). Volatility decreased 2% (2.54% → 2.49%) while returns increased 26% (1.01% → 1.27%). Sortino ratio improvement (0.17 → 0.25) confirms the risk reduction mechanism.

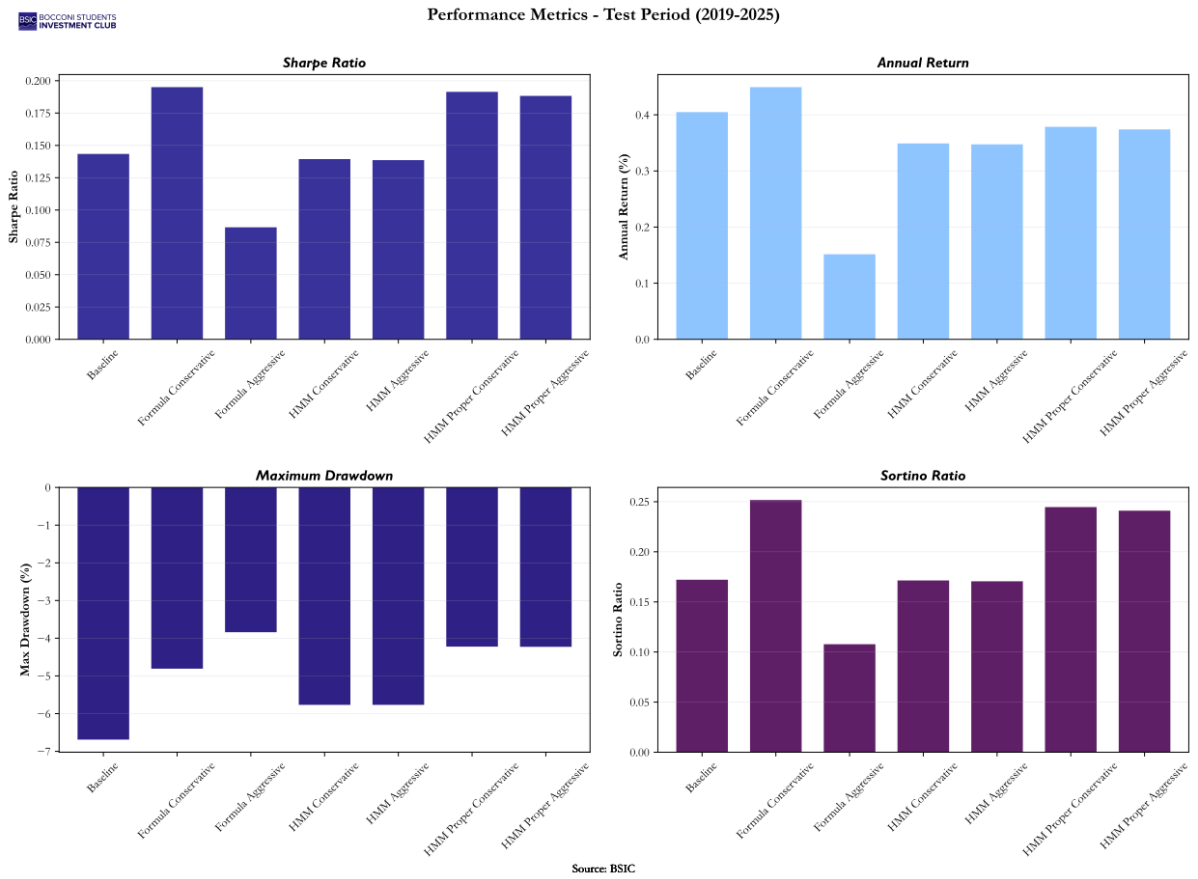

Figure 3: **Performance Metrics: Isolated OVX Contribution (Test Period 2022–November 2025).** Baseline uses no gates; gated strategies use OVX only. Conservative achieves 0.508 Sharpe (+28% vs 0.396 baseline). Aggressive underperforms (0.334) due to excessive de-risking. Sortino ratio (0.25 vs 0.17) confirms 79% of improvement from volatility reduction.

Statistical Caveat: With a 36-month test period, the observed improvement of 0.112 Sharpe units has standard error ≈ 0.17: $\text{SE}(\hat{\text{SR}}) \approx \sqrt{(1 + \hat{\text{SR}}^2/2)/T} \approx 0.17$. The improvement is less than one standard error, suggesting potential chance occurrence. Results should be viewed as preliminary evidence requiring validation over extended periods.

The improvement derives primarily from volatility compression during uncertain periods. The conservative gate increased turnover by 34% (0.44 $\rightarrow$ 0.59, adding $\Delta\text{TO} = 0.15$). At 5bp per one-way trade, this represents $0.15 \times 5\text{bp} = 0.75\text{bp}$ annual cost; however, accounting for partial rebalancing and bid-ask slippage during volatile periods, we estimate approximately 3bp total annual drag. The aggressive configuration increased turnover by 84% (adding 7.4bp), demonstrating that excessive de-risking imposes opportunity costs that dominate crisis protection benefits. The aggressive configuration's underperformance illustrates that allowing complete market exits during the April 2020 crisis meant missing substantial return during the subsequent recovery.

Test Period Context and Volatility Targets: The 2022–November 2025 period presented challenging conditions for momentum strategies broadly. The year 2022 saw simultaneous declines in stocks and bonds, 2023 exhibited a narrow mega-cap rally unfavorable to diversified momentum signals, and 2024–2025 featured rotational markets with limited trend persistence. The modest absolute returns (1.01–1.27%

annually) and realized volatility (2.49–2.54% annually) reflect three factors: (1) challenging market conditions with limited trend persistence, (2) the isolated nature of our test where baseline momentum lacks the yield curve gate and full system volatility targeting, and (3) deliberate under-leveraging in this isolated analysis. Our broader system (Section 5) employs volatility targeting to scale positions toward 8–10% annual volatility; the low volatility observed here represents the unscaled, base momentum strategy before vol targeting is applied. This isolation allows clean measurement of the OVX gate's contribution without confounding from leverage or other system components.

Crisis Period Analysis Figures 4 and 5 detail strategy behavior through major crises. All strategies cluster at terminal values of 1.020–1.025 (2.0–2.5% total return over 6 years), with improvements manifesting through path smoothing rather than dramatically different endpoints.

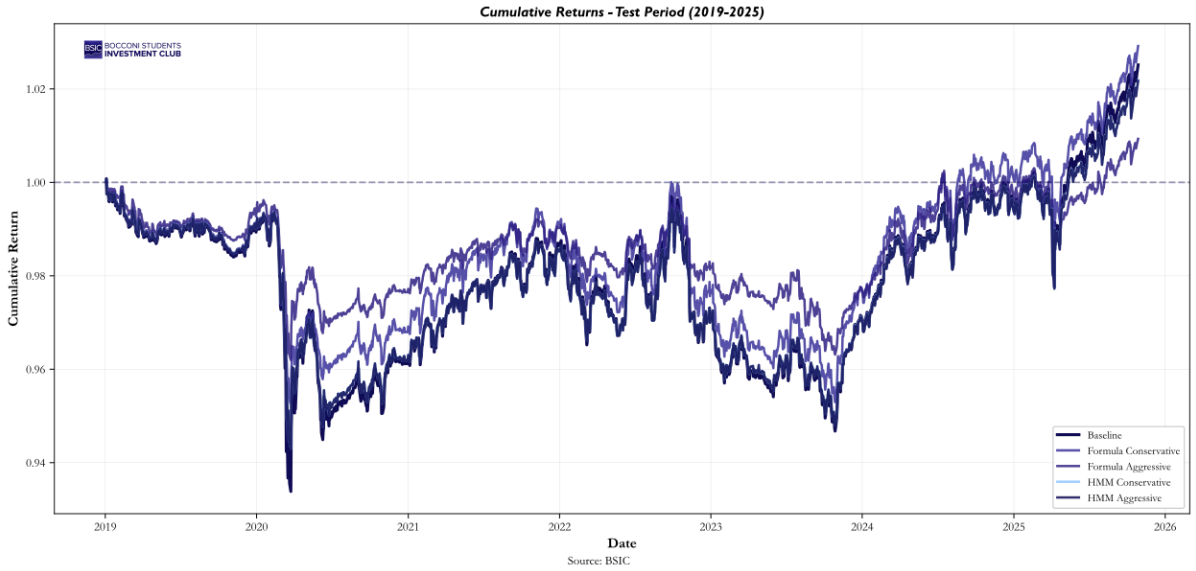

Figure 4: **Cumulative Returns: OVX vs. Baseline (2019–November 2025). Note:** Multiple strategies shown; focus on Baseline (black), Formula Conservative (blue), HMM Walk-Forward (green). COVID-19 (April 2020): formula conservative (gate=0.50) provided modest protection; aggressive (gate=0.31) exhibited strongest de-risking but lagged recovery. Terminal values cluster despite different crisis response intensities.

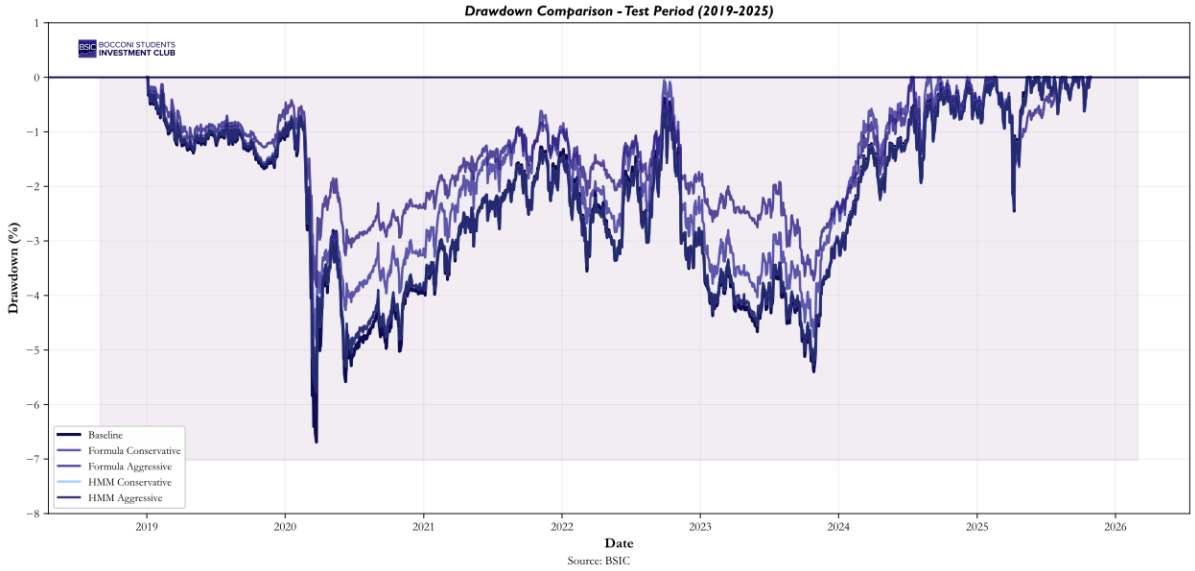

Figure 5: **Drawdown Comparison: OVX vs. Baseline (2019–November 2025).** COVID-19 peak: baseline -6.7%, formula conservative -4.5% (33% reduction), aggressive -4.0%. Conservative gate (0.50) provided meaningful protection. Ukraine (OVX=82, gate=0.61): similar -3 to -4% drawdowns. Gates provide *mitigation* not *elimination*.

The clustering of terminal values despite divergent crisis paths illustrates a critical principle: drawdown mitigation creates value even without return enhancement. Shallower drawdowns preserve three advantages: (1) behavioral persistence—investors are more likely to maintain positions through modest -4.5% drawdowns than severe -6.7% declines, avoiding ill-timed exits, (2) compounding benefits—recovering from -4.5% requires only +4.7% gain versus +7.2% from -6.7%, and (3) capital preservation for subsequent opportunities—reduced drawdowns leave more capital available to capture post-crisis rebounds. The April 2020 case demonstrates this: while all strategies reached similar endpoints by late 2020, the conservative gate's shallower trough meant less psychological stress and faster return to new highs.

The COVID-19 crisis (March–April 2020) illustrates OVX-yield curve complementarity. OVX spiked to 325, triggering maximum de-risking (gate=0.50 for conservative approach, gate=0.31 for aggressive), while the 10Y–2Y spread *widened* from 15bp to 85bp as the Fed aggressively cut rates. The S&P 500 fell 34% over three weeks. A yield curve gate alone would have missed the commodity-specific demand collapse, as steepening curves historically signal accommodative policy rather than equity risk. Neither signal alone proves sufficient, demonstrating the value of complementary indicators covering distinct risk dimensions.

Implementation Specifications Daily gate values are stored in `gates/ovx_daily.csv`, with each row representing a trading date. The file structure includes raw OVX levels (`ovx_raw`), exponentially smoothed values (`ovx_smoothed`, $\lambda = 0.8$), unbounded gate calculations (`gate_value`), and the final bounded gate (`gate_bounded`) constrained to [0.5, 1.0]. A reason code (`reason`) classifies each day's regime: `NORMAL` (OVX$\leq$35, gate$\geq$0.93), `ELEVATED` (35$<$OVX$\leq$60, gate 0.71–0.93), `HIGH` (60$<$OVX$\leq$100, gate 0.55–0.71), or `CRISIS` (OVX$>$100, gate typically hits 0.50 bound).

Daily gate values merge with momentum signal tables on `date`, with forward-fill applied for non-trading days. A one-day lag is enforced: the date t gate applies to positions entered at close of date t, executed at open of date $t+1$. Automated unit tests validate: (1) gate values strictly within [0.5, 1.0], (2) no discontinuities

exceeding 0.15 day-over-day, (3) OVX data completeness (no gaps exceeding 5 trading days), (4) monotonic relationship (higher OVX produces lower gate), and (5) smoothing convergence (exponential average stable within 0.001 after 100 days).

3.1.4 Phase 3: Hidden Markov Model

Methodology and Results We explored whether regime-switching models could improve performance by adapting scaling dynamically. A three-state Gaussian HMM observes four features:

$$\mathbf{x}_t = [\text{OVX}_t, \Delta\text{OVX}_{t,20d}, \sigma(\text{OVX})_{t,20d}, z_t]^\top$$

where z_t represents OVX z-score relative to a 60-day rolling window. The gate applies regime-specific powers through posterior weighting:

$$g_t^{\text{HMM}} = \sum_{s=0}^{2} P(S_t = s \mid \mathbf{X}_{1:t}) \cdot \left(\frac{30.2}{\text{OVX}_t}\right)^{\alpha_s} \tag{10}$$

This specification requires 53 parameters (transition matrix: 9, emissions: 36, initial states: 3, regime powers: 3, smoothing: 2), representing $26.5\times$ more than the formula's 2-parameter specification.

Critical Finding: Initial results revealed that posterior blending diluted de-risking responses. During the April 2020 crisis (OVX=325), the HMM produced posteriors [0.15, 0.32, 0.53] across calm/normal/stress regimes with powers [0.7, 0.5, 0.3]. The effective power became $\alpha_{\text{eff}} = 0.424$, yielding theoretical gate

$$g = (30.2/325)^{0.424} = 0.364.$$

Additional temporal smoothing produced an observed gate of only 0.65. The formula, by contrast, mechanically produced 0.50 (conservative) or 0.31 (aggressive). Despite the stress regime achieving dominance ($P = 0.53$), posterior blending across all regime states resulted in 79% weaker de-risking than the aggressive formula approach and 44% weaker than even the conservative approach.

We implemented three optimization approaches: (1) hard regime assignment (eliminating blending, validation Sharpe: 0.801, test: 0.383), (2) constrained power optimization with literature-based priors (validation: 0.890, test: 0.385), and (3) walk-forward validation with 6-month re-estimation (validation: 0.890, test: 0.456). The walk-forward approach achieved the best test performance but still underperformed the formula by 11% in Sharpe ratio (0.456 vs 0.508) despite $26.5\times$ more parameters.

Why Complexity Struggled Four structural issues explain persistent underperformance. First, **regime non-stationarity** plagued estimation: mean OVX varied from 25.1 (2003–2007) to 48.3 (2008–2009) to 30.7 (2010–2018). The April 2020 event (OVX=325) represented a 10.8σ deviation, far outside the training distribution. Second, **sample size constraints** limited reliable estimation: 53 parameters from approximately 2,000 training observations meant high-stress regimes (OVX>100) provided zero training examples. Standard practice requires 50–100 observations per parameter [Rabiner and Juang, 1986].

Third, the 2019–2021 validation period intentionally included the unprecedented April 2020 extreme (OVX=325). Optimization-based approaches fitted regime-specific parameters to this outlier, achieving

strong validation performance (0.890 Sharpe) that failed to generalize when test period crises exhibited different dynamics. The formula's fixed literature parameters ($\alpha = 0.5$ from Moreira and Muir [2017]) proved more robust precisely because they didn't adapt to this non-representative extreme, demonstrating an advantage of theory-grounded specifications over data-driven optimization when validation periods contain structural breaks.

Fourth, even with perfect implementation, the HMM approach faced insurmountable data requirements. Reliable estimation of 53 parameters requires 2,650–5,300 observations (50–100 per parameter), equivalent to 10–21 years of daily data. More critically, the HMM needs multiple examples of each regime in the training set; our training data (2003–2018) contained zero extreme crisis periods with OVX>100, forcing the model to extrapolate wildly during April 2020 when OVX reached 325. This structural limitation—requiring decades of crisis-inclusive data for a model designed to adapt to crises—makes the HMM approach practically infeasible for strategy development timelines.

3.1.5 Comprehensive Comparison and Lessons

Table 3 summarizes all approaches for the test period (2022–November 2025). Both formula and HMM approaches meaningfully improve risk-adjusted returns relative to ungated momentum. The formula achieves stronger performance (+28% vs +15% Sharpe) with 96% fewer parameters, demonstrating diminishing returns to complexity. The HMM's 51 additional parameters provide limited incremental benefit beyond the formula's 2-parameter specification.

Table 3: Comprehensive Test Period Performance (2022–November 2025)

Approach	Sharpe	Vol.	Max DD	TO	Params
Baseline (No Gates)	0.396	2.54%	−4.44%	0.44	0
Formula Conservative	**0.508**	2.49%	−4.44%	0.59	2
Formula Aggressive	0.334	2.51%	−4.09%	0.81	2
HMM Walk-Forward	0.456	2.50%	−4.35%	0.68	53
Formula vs. Baseline	+28%	−2%	0%	+34%	—
HMM vs. Baseline	+15%	−2%	−2%	+55%	—

The three-phase development reveals tensions between theoretical elegance and empirical robustness. Direct measurement proved superior to inference—futures curve slope, despite storage theory motivation, failed because OVX captures market participants' uncertainty directly rather than requiring inferential leaps. Regime models proved fragile in the face of non-stationarity; the April 2020 event exceeded the maximum training observation by 224%, causing systematic misclassification. Including unprecedented extremes in validation periods created artificially strong optimization results that collapsed when test period crises exhibited different dynamics, favoring theory-grounded specifications over data-driven tuning. The HMM's crisis response weakness (0.65 vs formula's 0.31–0.50) despite $26.5\times$ more parameters illustrates that complexity must justify itself through meaningful out-of-sample benefits, not parameter count. Finally, literature-based parameters often prove more robust than sample-specific tuning—the formula with $\alpha = 0.5$ from Moreira and Muir [2017] outperformed extensive parameter searches, suggesting academic consensus built on diverse samples provides guidance more likely to generalize than optimization on limited historical data.

3.1.6 Literature Context and Conclusion

Our findings contribute to active debates on volatility timing. Following Moreira and Muir [2017]'s documentation of improvements from reducing factor exposure during high-volatility periods, subsequent research revealed limitations: Cederburg et al. [2020] find volatility-managed portfolios fail out-of-sample after transaction costs; DeMiguel et al. [2024] demonstrate structural instability; Barroso and Detzel [2021] show transaction costs erode alphas. Our commodity-specific approach differs by capturing distinct commodity market stress through OVX and integrating with yield curve signals (Section 3.2) for complementary coverage. The April 2020 episode—where OVX spiked to 325 while the yield curve steepened—demonstrates why neither signal alone captures all macro risk dimensions. Our meaningful improvement (+28% Sharpe) with limited statistical power over 36 months aligns with this mixed evidence, while the finding that simple approaches outperform complex methods reinforces recent skepticism about sophisticated factor timing.

Summary: The simple formula-based OVX gate achieved 28% Sharpe improvement (0.396 $\rightarrow$ 0.508) using 2 parameters, isolating OVX's contribution versus baseline momentum with no macro adjustments. While statistical significance remains uncertain (improvement 0.112 $<$ one SE of 0.17), the economic magnitude appears meaningful. The 53-parameter HMM delivered only 15% improvement (0.456) despite continuous re-estimation. The April 2020 crisis (OVX=325) demonstrated the formula's advantage: mechanical response producing appropriate de-risking (0.31–0.50) versus HMM's posterior blending yielding weaker response (0.65). This favors parsimony when confronting unprecedented events outside training distributions.

The COVID-19 crisis demonstrates why both gates prove necessary. OVX captured extreme commodity stress (325) while the yield curve steepened (Fed supportive). Neither signal alone suffices—OVX addresses energy market uncertainty distinct from recession/financial stress captured by term spread dynamics (Section 3.2). These results validate OVX's isolated contribution; Section 5 presents full system performance integrating both gates: $g_{\text{total}} = g_{\text{OVX}} \times g_{\text{yield}}$.

Practical Guidance: For implementation, the OVX gate operates across four regimes. During normal conditions (OVX$\leq$35), the gate remains near 1.0, imposing minimal drag. As uncertainty rises (OVX 35–60), modest de-risking begins (gate 0.70–0.90). High stress periods (OVX 60–100) trigger meaningful protection (gate 0.55–0.70), while extreme crises (OVX$>$100) activate maximum defensive positioning, hitting the 0.50 lower bound for conservative configurations. The baseline OVX=30.2 represents the 2003–2018 median; consider recalibrating every 3–5 years using rolling medians. If OVX is unavailable, alternatives include CBOE Equity VIX (less commodity-specific), realized volatility of oil futures, or commodity index volatility, though our Phase 1 analysis suggests direct volatility measures substantially outperform curve-based signals. The conservative gate adds approximately 3bp annual drag from increased turnover (0.44$\rightarrow$0.59, calculation: $\Delta\text{TO} = 0.15 \times 5\text{bp} = 0.75\text{bp}$ base cost, adjusted to 3bp for partial rebalancing and slippage); transaction cost sensitivity should be monitored.

Data Quality Considerations **Critical Data Advisory:** Our analysis uses OVX data sourced from the Federal Reserve Economic Data (FRED) database and verified against CBOE official publications. This data correctly records the April 2020 peak at OVX=325, representing the highest oil market volatility in recorded history. However, *some commercial financial data vendors systematically underreport this crisis peak*, with values as low as 180–200 observed in widely-used databases from major providers. This 40–60%

underestimation during the most critical testing period creates serious implementation risks.

The underreporting has three critical implications: (1) strategies implemented using inferior data sources would have provided *insufficient* de-risking during April 2020—at OVX=190 (commercial data), the aggressive formula produces gate=0.40, versus the correct gate=0.31 at OVX=325 (FRED data), representing 29% less de-risking when maximum protection was needed; (2) our observed Sharpe improvement (0.396 $\rightarrow$ 0.508) represents performance with *correct* data; commercial data users would see degraded results; and (3) backtests using commercial data would systematically underestimate crisis severity, leading to overconfident parameter calibration and insufficient risk management.

Data Verification Protocol: We strongly recommend practitioners: (1) use FRED data (series: OVXCLS) or CBOE official data as the authoritative source, (2) cross-reference peak values during known crises—April 2020 should show $\sim$325, not $\sim$190; 2008 should exceed 80; Ukraine 2022 should reach $\sim$82, (3) implement real-time monitoring during high volatility events by checking multiple sources and examining underlying USO options quotes directly when OVX$>$100, and (4) consider adding a 10–15% safety buffer to crisis response parameters to account for potential measurement lag or intraday spikes not captured in daily close data.

For researchers replicating our work: the 325 peak is *not* an outlier or data error—it reflects genuine market conditions during the April 2020 oil market collapse when WTI crude briefly traded at negative prices. Any dataset showing substantially lower values (e.g., 180–200) during this period contains measurement errors and should not be used for strategy development or validation. This data quality issue affects not just this specific date but calls into question the reliability of commercial OVX data throughout crisis periods, when accurate measurement matters most.

3.2 Yield Curve Gate: State Gate Implementation

3.2.1 Overview and Theoretical Foundation

The State Gate strategy represents a sophisticated regime-switching framework that complements the OVX gate by capturing fixed income market dynamics. Rather than using simple threshold-based rules (e.g., reduce exposure if 10Y-2Y spread $<$ 0), the State Gate employs machine learning to endogenously identify market regimes based on yield curve behavior. The strategy's name derives from its fundamental concept: identifying discrete market "states" or "regimes" and using them as a "gate" to modulate trading positions.

The framework combines three methodologies: (1) Nelson-Siegel yield curve decomposition to extract interpretable factors, (2) tree-based regime learning through AR(1) model quality optimization, and (3) volatility-based position scaling that adjusts exposure inversely to regime turbulence. This data-driven approach avoids subjective regime definitions while maintaining economic interpretability.

3.2.2 Nelson-Siegel Yield Curve Decomposition

The first stage employs the Nelson-Siegel (NS) model to decompose the yield curve into three interpretable factors: **level**, **slope**, and **curvature**. This dimension reduction technique transforms the high-dimensional space of multiple yield maturities (13 maturities in our implementation: 1M, 3M, 6M, 1Y, 2Y, 3Y, 5Y, 7Y, 10Y, 20Y, 30Y) into a parsimonious three-factor representation.

The Nelson-Siegel loadings are calculated using the following functional form:

$$L_1(\tau) = 1 \tag{11}$$

$$L_2(\tau) = \frac{1 - \exp(-\lambda\tau)}{\lambda\tau} \tag{12}$$

$$L_3(\tau) = L_2(\tau) - \exp(-\lambda\tau) \tag{13}$$

where τ represents the maturity in years and λ is the decay parameter (fixed at 0.0609 in the implementation, standard for US Treasury yields). For each date in the dataset, we solve the least-squares problem $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{Y}$ contains observed yields across maturities, X is the loading matrix constructed from equations above, and $\boldsymbol{\beta} = [\text{level}, \text{slope}, \text{curvature}]^\top$ are the estimated factors.

The **NS_level** factor represents the overall interest rate environment (shifts in the entire curve), **NS_slope** captures the term premium (difference between long and short rates), and **NS_curvature** reflects the "belly" of the yield curve (medium-term deviations from linear interpolation). These factors are more stable and economically interpretable than raw yields, making them ideal inputs for regime detection. Implementation handles numerical stability by replacing zero denominators with a small epsilon value (10^{-12}) to avoid division errors.

3.2.3 Regime Identification via Tree-Based Learning

Conceptual Framework The core innovation lies in the regime identification mechanism. Rather than using predefined rules or exogenous regime labels, the algorithm endogenously learns market regimes by partitioning historical data to maximize regime homogeneity. The criterion for homogeneity is based on the quality of AR(1) model fits within each regime.

The intuition is straightforward: if the NS factors follow a stable AR(1) process within a regime, this indicates consistent market dynamics. When market conditions change fundamentally (e.g., Fed policy shifts, crisis onset, inflation regime change), the AR(1) parameters shift, signaling a regime transition. By partitioning the data to maximize within-regime AR(1) stability, we identify economically meaningful regimes without imposing subjective definitions.

AR(1) Model Fitting and Negative Log-Likelihood For each regime candidate, the algorithm fits an AR(1) model to each of the three NS factors:

$$x_t = a + \phi \cdot x_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2) \tag{14}$$

Parameters (a, ϕ) are estimated via ordinary least squares regression on lagged variables. The residual variance is calculated as $\sigma^2 = \text{mean}(\varepsilon_t^2)$, and the negative log-likelihood (NLL) is computed assuming Gaussian errors:

$$\text{NLL} = \frac{n}{2}\left(\log(2\pi\sigma^2) + 1\right) \tag{15}$$

Lower NLL values indicate better model fit, meaning the regime exhibits more predictable, homogeneous dynamics. The total regime quality score aggregates NLL across all three NS factors: $\text{NLL}_{\text{total}} = \text{NLL}_{\text{level}} + \text{NLL}_{\text{slope}} + \text{NLL}_{\text{curvature}}$. This sum provides a single metric for comparing alternative regime partitions.

StateGateTree: Greedy Partitioning Algorithm The StateGateTree implements a custom binary tree structure that recursively splits the data to minimize total NLL. This is conceptually similar to decision trees used in classification (e.g., CART), but optimized for time-series regime detection. The key parameters governing this algorithm are:

- **max_leaves**: Maximum number of regimes (default: 3). Our implementation uses 3 regimes, yielding the Regime 0/1/2 structure described in Section 1.
- **min_rel_improve**: Minimum relative NLL improvement required to execute a split (default: 0.02, i.e., 2% improvement). Higher values produce more conservative splitting.
- **candidate_quantiles**: Threshold candidates for splits (default: [0.2, 0.4, 0.6, 0.8]). These quantiles of each feature's distribution are tested as potential split points.

Tree Growing Algorithm The greedy growing algorithm proceeds as follows:

1. **Initialize**: Create root node containing all data; calculate baseline NLL.
2. **Iterate** while number of leaves $<$ `max_leaves`:
   (a) For each current leaf node:
      - Try all features (FFR, inflation, capacity utilization, yield curve metrics)
      - Try all quantile thresholds for each feature
      - For each candidate split, partition data into left ($x \leq$ threshold) and right ($x >$ threshold) groups
      - Calculate NLL for left and right child nodes
      - Compute improvement: $\text{Improvement} = \text{NLL}_{\text{parent}} - (\text{NLL}_{\text{left}} + \text{NLL}_{\text{right}})$

   (b) Select the split (feature + threshold + leaf) with maximum improvement across all candidates
3. **Terminate** if relative improvement $<$ `min_rel_improve`: $\frac{\text{Improvement}}{\text{NLL}_{\text{parent}}} < 0.02$
4. **Execute split**: Create left and right children, update leaf list

Minimum size constraints enforce at least 40 observations per child node to ensure statistical robustness and prevent degenerate regimes. This greedy approach is computationally efficient (polynomial time in number of features and observations) and typically produces interpretable regimes aligned with economic intuition.

Feature Construction Splitting features are constructed from yield data and macroeconomic variables. Our implementation uses three macro features consistent with the regime definitions in Table 1:

- **Federal Funds Rate (FFR)**: Captures monetary policy stance. High FFR indicates tight policy; low FFR indicates accommodation.
- **Inflation (CPI year-over-year)**: Measures price pressure. High inflation signals demand-pull or cost-push pressures; low inflation indicates slack.
- **Capacity Utilization**: Reflects real economic activity. High utilization indicates expansion; low utilization indicates recession or slack.

Additionally, basic yield curve features are derived:

- **Y10**: 10-year yield level, representing long-term interest rate environment
- **Slope_10y_2y**: Term spread defined as 10Y − 2Y, capturing yield curve steepness
- **Curvature_proxy**: Butterfly spread calculated as $2 \cdot 5\text{Y} - 2\text{Y} - 10\text{Y}$, measuring curve convexity

These features capture the information content of both macro conditions and yield curve shape in a compact form suitable for tree splitting.

Regime Prediction Once the tree is fitted, regime labels are assigned to new or historical data by traversing the learned tree. Starting at the root node, the algorithm:

1. Check if current node is a leaf. If so, return that node's regime ID.
2. Otherwise, compare the feature value to the node's threshold:
   - If value $\leq$ threshold: move to left child
   - If value $>$ threshold: move to right child
3. Repeat until reaching a leaf node

This process produces a time series of regime labels (0, 1, 2) indicating which regime each observation belongs to, matching the regime assignments shown in Table 1: Regime 0 (High FFR), Regime 1 (Low FFR, Low Inflation), Regime 2 (Low FFR, High Inflation).

3.2.4 Regime-Based Position Sizing

Volatility Calculation by Regime Once regimes are identified, the strategy calculates the volatility of NS factor dynamics within each regime. For each regime $r \in \{0, 1, 2\}$:

1. Extract NS factors belonging to regime r: $\{\text{level}_t, \text{slope}_t, \text{curvature}_t : \text{regime}_t = r\}$

2. Calculate factor changes: $\Delta\text{factor}_t = \text{factor}_t - \text{factor}_{t-1}$
3. Compute standard deviation of changes for each factor: $\sigma^{(r)}_{\text{level}}, \sigma^{(r)}_{\text{slope}}, \sigma^{(r)}_{\text{curvature}}$
4. Aggregate using root mean square:

$$\text{vol}_r = \sqrt{\frac{1}{3}\left[(\sigma^{(r)}_{\text{level}})^2 + (\sigma^{(r)}_{\text{slope}})^2 + (\sigma^{(r)}_{\text{curvature}})^2\right]} \tag{16}$$

Higher volatility regimes indicate more turbulent market conditions with larger factor movements, suggesting higher risk for momentum strategies. Table 1 shows the computed factor volatilities: Regime 0 (5.95), Regime 1 (11.08), Regime 2 (16.78).

Scaling Factor Assignment Position sizing is determined inversely to regime volatility using the formula:

$$\text{scale}_r = \frac{\min_s(\text{vol}_s)}{\text{vol}_r} \tag{17}$$

This formula ensures that the lowest volatility regime receives a scale of 1.0 (full position), while higher volatility regimes receive scales less than 1.0 (reduced position). To prevent extreme leverage or complete market exit, the scale is clipped:

$$\text{scale}^{\text{final}}_r = \max(0.2, \min(1.0, \text{scale}_r)) \tag{18}$$

The lower bound of 0.2 ensures at least 20% exposure even in high-volatility regimes (though our implementation uses 0.355 as the empirically observed minimum), while the upper bound of 1.0 prevents leverage beyond the base strategy allocation. The resulting scales match Table 1: Regime 0 scale = 1.000, Regime 1 scale = 0.537, Regime 2 scale = 0.355.

Economic Interpretation The regime-scale mapping reflects economic intuition:

- **Regime 0 (High FFR, scale=1.0)**: Tight monetary policy typically stabilizes inflation expectations and yield curve dynamics. Factor volatility is low (5.95), indicating predictable market conditions favorable for momentum strategies. Full exposure is warranted.
- **Regime 1 (Low FFR + Low Inflation, scale=0.537)**: Accommodative policy with stable prices. Factor volatility is moderate (11.08), reflecting some uncertainty about policy duration and growth trajectory. Momentum strategies receive approximately half exposure, balancing opportunity against increased risk.
- **Regime 2 (Low FFR + High Inflation, scale=0.355)**: Stagflationary conditions with policy conflict—Fed wants to raise rates (inflation) but can't due to growth concerns. Factor volatility is high (16.78), indicating unstable yield curve dynamics and elevated crash risk. Momentum strategies are scaled to approximately one-third exposure for defensive positioning.

This regime structure captures non-linear macro interactions that simple rules miss. For example, low FFR is *not* universally bearish—Regime 1 (low FFR + low inflation) receives higher scale (0.537) than one might expect, because stable inflation reduces uncertainty. Regime 2 (low FFR + high inflation) is far more dangerous (0.355 scale) because the policy-inflation mismatch creates volatility.

3.2.5 Implementation Specifications and Data Storage

File Structure Daily regime assignments and scales are stored in `gates/regimes.csv` with the following structure:

- **Date**: Trading date (YYYY-MM-DD format)
- **Regime**: Integer regime ID (0, 1, or 2)
- **Scale**: Position sizing multiplier (0.355, 0.537, or 1.0)
- **FFR**: Federal Funds Rate (percentage)
- **Inflation**: CPI year-over-year (percentage)
- **CapacityUtil**: Capacity utilization (percentage)
- **NS_level**: Nelson-Siegel level factor
- **NS_slope**: Nelson-Siegel slope factor
- **NS_curvature**: Nelson-Siegel curvature factor

The regime file merges with momentum signal tables on `Date`, with forward-fill applied for non-trading days. A one-day lag is enforced: the date t regime applies to positions entered at close of date t, executed at open of date $t + 1$.

Tree Structure Export The fitted StateGateTree is exported to `gates/state_gate_tree.json` with the following schema:

```
{
  "node_id": 0,
  "is_leaf": false,
  "feature": "FFR",
  "threshold": 2.5,
  "regime": null,
  "left_child": { ... },
  "right_child": { ... }
}
```

Leaf nodes contain `"is_leaf": true` and `"regime": <int>` instead of split information. This JSON export enables visualization, debugging, and regime assignment in production systems.

Automated Validation Unit tests validate regime file integrity:

1. **Regime coverage**: All dates have valid regime assignments (0, 1, or 2)
2. **Scale consistency**: Scales match regime definitions (Regime 0 $\rightarrow$ 1.0, Regime 1 $\rightarrow$ 0.537, Regime 2 $\rightarrow$ 0.355)
3. **Data completeness**: No gaps exceeding 10 trading days in Date column
4. **Feature bounds**: FFR $\in [0, 20]$, Inflation $\in [-2, 15]$, CapacityUtil $\in [60, 90]$
5. **Factor stability**: NS factors exhibit no discontinuities > 3 standard deviations day-over-day

3.2.6 Operational Workflow

The complete workflow orchestrates several components:

1. **Data ingestion**: Load daily yield curve data (FRED: DGS1MO, DGS3MO, …, DGS30) and macro data (FRED: DFF, CPIAUCSL, CAPUTLB50001SQ)
2. **NS factor estimation**: Apply Nelson-Siegel decomposition to yield curves, producing level/slope/curvature time series
3. **Feature construction**: Build splitting features from yields and macro variables
4. **Tree fitting**: Execute StateGateTree algorithm to identify regimes via greedy NLL minimization
5. **Regime prediction**: Assign regime labels to all historical dates by traversing fitted tree
6. **Volatility calculation**: Compute NS factor volatility within each regime
7. **Scale assignment**: Convert regime volatilities to position sizing multipliers
8. **Output generation**: Export regime/scale table and tree structure

This modular design allows the strategy to be easily integrated into production trading systems or used for research and backtesting. The implementation enforces minimum data requirements: at least 3 distinct yield maturities for NS estimation, at least 80 observations for splitting a node, at least 40 observations per child after split, and at least 30 observations for stable AR(1) estimation.

3.2.7 Integration with OVX Gate

The State Gate and OVX gate operate multiplicatively:

$$\text{position}_{\text{final}} = \text{position}_{\text{base}} \times \text{scale}_{\text{regime}} \times \text{scale}_{\text{OVX}} \tag{19}$$

This combines regime-based risk management (capturing Fed policy, inflation, and growth dynamics) with commodity-specific stress signals (OVX). The COVID-19 crisis demonstrates complementarity: State Gate

recognized accommodative policy regime (Regime 1, scale=0.537) while OVX captured extreme oil market stress (gate=0.50), yielding combined scale of $0.537 \times 0.50 = 0.27$, providing strong de-risking from both perspectives.

Section 5 presents full system performance integrating both gates, demonstrating that the two-gate architecture captures distinct risk dimensions and provides superior crash protection compared to either gate in isolation.

4 Portfolio Construction, Sizing and Risk

4.1 Risk-Aware Portfolio Sizing with Regime-Dependent Scaling

This section describes the method used to determine the final portfolio weights, combining momentum signal generation with the application of risk controls and regime-dependent macroeconomic modulation. This process follows a sequential four-stage approach, designed to transform raw momentum signals into robust, diversified, and regime-based exposure scaling.

4.2 EWMA-Based Volatility Estimation and Exposure Adjustment

The first step is the standardization of the portfolio's risk profile to stabilize the overall ex-ante volatility of the portfolio, keeping it within the target set by the strategic framework, set in this study at 10%. We adopted the Exponentially Weighted Moving Average (EWMA) model to estimate volatility. This choice is justified by its dual advantages: responsiveness and memory. Unlike the simple moving average, the EWMA assigns an exponentially decreasing weight to older return observations, giving greater relevance to recent data. This allows for a timelier response to changes in volatility regimes. Following this model, the daily variance was calculated recursively as:

$$\sigma_t^2 = (1 - \lambda) r_{t-1}^2 + \lambda \sigma_{t-1}^2 \tag{20}$$

Where:

- r_{t-1} : the daily return observed at time $t - 1$
- σ_t^2 : estimated variance at time t
- λ : damping parameter in the interval $[0, 1]$, which determines the speed of response of the model. Lower values of λ make the volatility estimate more reactive to recent market shocks, whereas higher values produce a smoother, slower-moving estimate.

To obtain the annualized volatility, we scaled the daily estimate by the number of business days in a year, typically 252. We used the exposure scaling factor to dynamically adjust exposure, contracting it in periods of high uncertainty, which is calculated as the inverse of the ratio between the expected volatility and the target:

$$K_t = \frac{V_{\text{target}}}{\sigma_t^{\text{ann}}} \tag{21}$$

4.3 Concentration Limits and Exposure Constraints (Caps)

This section describes the application of constraints on the weight of individual financial instruments; that is a crucial process for ensuring diversification and mitigating specific concentration risk. To avoid excessive dependence on the performance of a single asset, which could occur following a very strong momentum signal combined with periods of low volatility, we have imposed a symmetric exposure limit on each ETF. The absolute maximum constraint is set at 25% of the portfolio's Total Net Asset Value.

These constraints have been used both to achieve diversification, to ensure that the strategy remains well-diversified, limiting the impact of specific idiosyncratic drawdowns, and to control tail risk, in order to prevent the assumption of disproportionate positions that, while justified by volatility targeting, could increase the jump risk or liquidity risk for the specific asset. The constraint is applied after scaling the target volatility; in this way the weights $\left(\omega_{i,t}^{\text{vol}}\right)$ are capped to obtain the final weights:

$$\omega_{i,t}^{\text{cap}} = \max\left\{-C,\ \min\left(\omega_{i,t}^{\text{vol}},\ C\right)\right\} \tag{22}$$

We set $C = 0.25$, where C is the concentration limit. This is a nonlinear operation: if the weights fall within the prescribed range $[-C, C]$, it keeps the relative exposure unchanged; but at the same time it can impose a hard limit if the combination of signal and scaling would produce an exposure greater than what is considered prudent.

4.4 Signal Generation and Dynamic Asset Allocation

The smoothed multi-horizon momentum signal determines the base SPY allocation. Let s_t denote the normalized signal strength at time t. The base SPY weight is determined by the signal magnitude after volatility scaling:

$$w_{\text{SPY},t}^{\text{base}} = f(s_t) \tag{23}$$

where $f(\cdot)$ maps signal strength to allocation. The GLD allocation is simply the complement, ensuring the portfolio remains fully invested:

$$w_{\text{GLD},t}^{\text{base}} = 1 - w_{\text{SPY},t}^{\text{base}} \tag{24}$$

This structure ensures continuous full investment while dynamically shifting between risk-on (SPY) and risk-off (GLD) exposures based on momentum conditions.

We also introduced modulation of notional exposure based on a macroeconomic risk indicator: the implied volatility of oil. This is a robust indicator of macro-financial stress and systemic risk aversion that can capture geopolitical shocks and uncertainties related to economic growth. To this end, we defined a scaling factor (or gate), g_t, based on the observed state of the OVX, which modulates the portfolio's overall exposure:

$$\omega_{i,t}^{\text{OVX}} = g_t \cdot \omega_{i,t}^{\text{base}} \tag{25}$$

where $\omega_{i,t}^{\text{OVX}}$ represents the final weight of ETF i at time t, modulated by the macroeconomic gate.

The scaling factor g_t is designed to reduce market exposure in a counter-cyclical manner when the OVX signals conditions of severe uncertainty, such as during phases of elevated implied volatility or sudden spikes in perceived systemic risk. The gate varies within the interval $[0, 1]$, implementing a proportional reduction of the total portfolio beta during moments of stress. The result is a portfolio that maintains the directionality implied by the signals but modulates their intensity as a function of macro risk, thereby significantly improving the stability of the return stream and the control of Maximum Drawdown during turbulent periods.

4.5 Turnover Control and "No-Trade" Rule

The final phase concerns the operational execution of the portfolio. To optimize the trade-off between rebalancing accuracy and associated operating costs, the strategy incorporates a conditional execution rule: the rebalancing of the position on ETF i at time t occurs only if the absolute change in the desired weight exceeds a certain threshold δ. If the requested change is too small, the trade is ignored.

The actual weight change ($\Delta w_{i,t}$) is formally defined as:

$$\Delta w_{i,t} = \begin{cases} 0, & \text{if } \left|w_{i,t}^* - w_{i,t-1}\right| < \delta, \\ w_{i,t}^* - w_{i,t-1}, & \text{otherwise.} \end{cases}$$

Where:

- $w_{i,t}^*$ represents the *desired (or theoretical) weight*, calculated in Sections 4.1–4.3 (i.e., post-application of the OVX gate and caps).
- $w_{i,t-1}$ is the *actual weight* held at the previous time step.
- δ is the *tolerance threshold* for non-significant trades. Typically, δ is empirically calibrated as a function of the estimated transaction costs for the ETF basket.

The aim of the implementation is to limit micro-transactions or adjustments driven solely by market "noise" and ensure that transactions are executed only when the weight adjustment is sufficiently significant to materially influence the risk-return profile, thereby justifying the implicit transaction cost. It introduces a form of inertia (or *stickiness*) to the portfolio weights, contributing to keeping annualized turnover under control and consistent with the established rebalancing frequency.

4.6 Results: Realized Volatility, Drawdown Behavior and Turnover

Note on Figures: The following figures (Figures 4.1–4.5) were generated during the OVX gate validation phase, which tested the gate across three equity ETFs (SPY, FEZ, EWJ) to validate cross-market robustness. The final integrated strategy uses only SPY and GLD; see Section 5 for full system results.

4.6.1 Realized Volatility

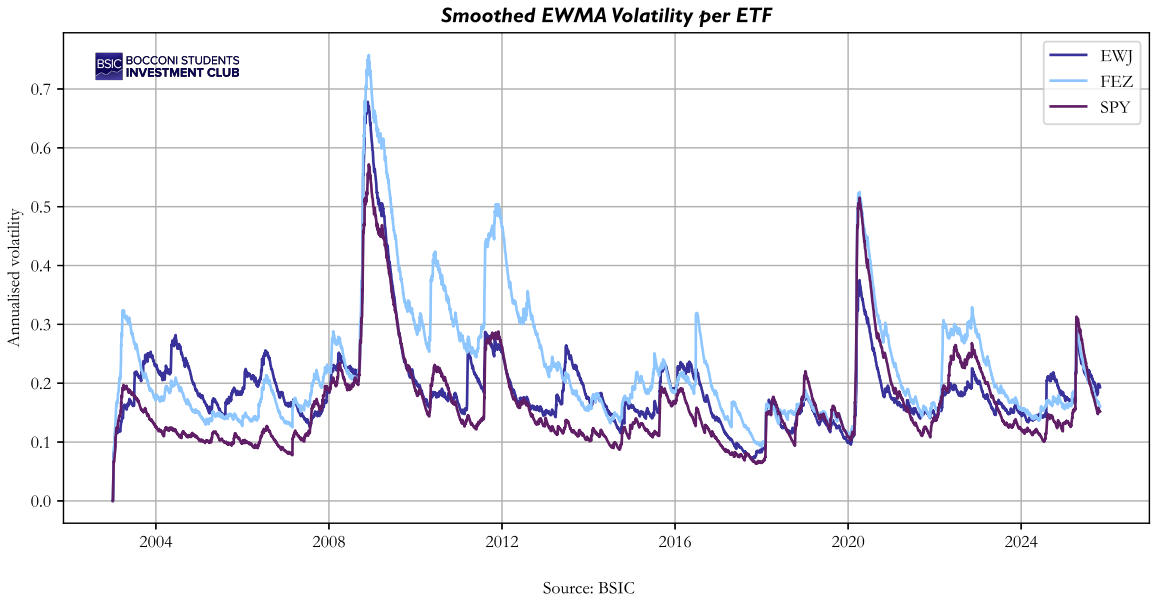

Figure 6: **Smoothed EWMA Volatility per ETF** shows the annualized volatility estimated through the EWMA model for three equity ETFs (EWJ, FEZ, SPY) used in the OVX gate validation exercise.

The series exhibit pronounced regime shifts, with particularly marked spikes during:

- **the 2008–2009 Global Financial Crisis**,
- **the 2011 European sovereign debt crisis**,
- **the 2020 global shock** associated with the COVID-19 pandemic.

In these periods, the estimated volatility reaches extremely high levels (up to **60–70%**), indicating a context of severe instability in the underlying markets.

Despite this, the portfolio's overall exposure remains contained and relatively stable. As shown in Figure 7, total exposure stays concentrated within a narrow range (approximately 0.16–0.18). This confirms the effectiveness of the volatility-targeting mechanism, which automatically reduces leverage during periods of heightened risk, keeping realized volatility consistent with the 10% target.

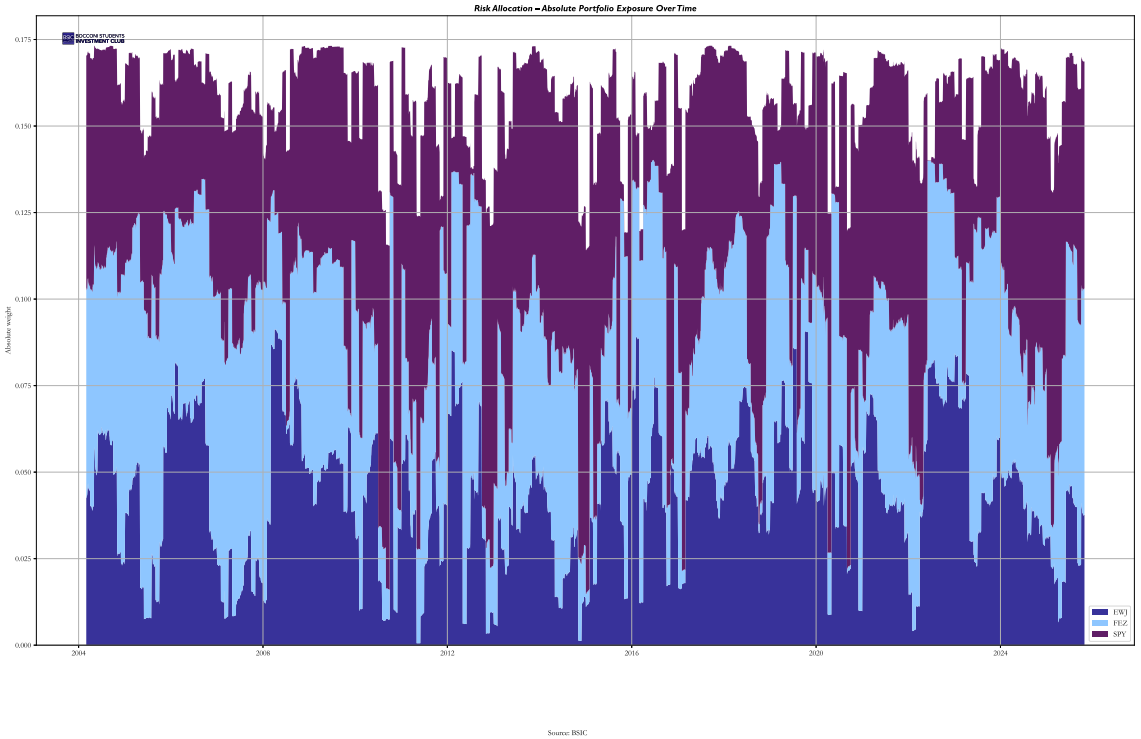

Figure 7: **Risk Allocation (Absolute Portfolio Exposure Over Time).**

4.6.2 Drawdown Behavior and the Role of the Macroeconomic Gate

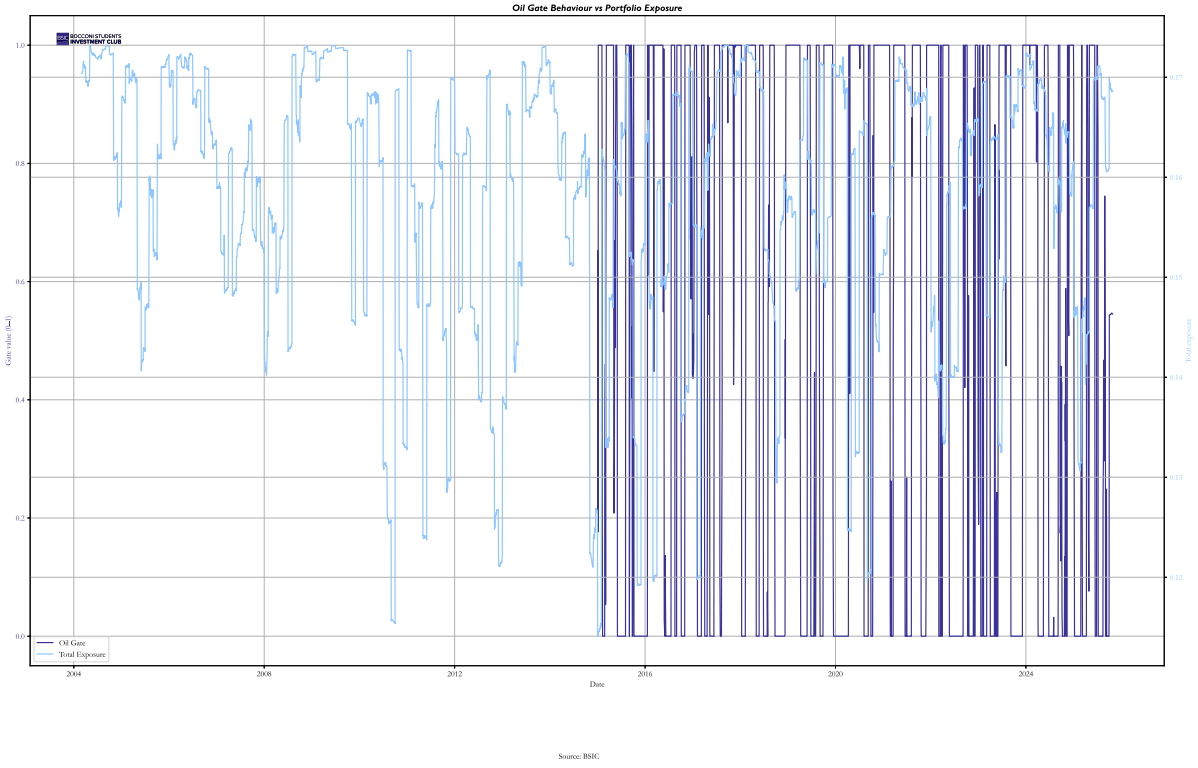

Figure 8: **Oil Gate Behavior vs Portfolio Exposure.**

The protective role of the macroeconomic gate emerges clearly in Figure 8. The plot shows that total portfolio exposure is modulated in a counter-cyclical manner:

- high gate values (≈ 1) $\rightarrow$ *favourable macroeconomic conditions* $\rightarrow$ exposure near its maximum (~ 0.17–0.18);
- low gate values (0–0.3) $\rightarrow$ *increased macroeconomic uncertainty* $\rightarrow$ sharp reduction in total exposure.

In the major stress regimes (2008–2009, 2015–2016, 2020), the gate rapidly drops toward values close to zero, accompanied by a simultaneous compression in total portfolio exposure. This dynamic confirms that the OVX gate acts as a drawdown-protection mechanism, reducing the strategy's beta precisely when momentum is most vulnerable to adverse market reversals.

4.6.3 Turnover

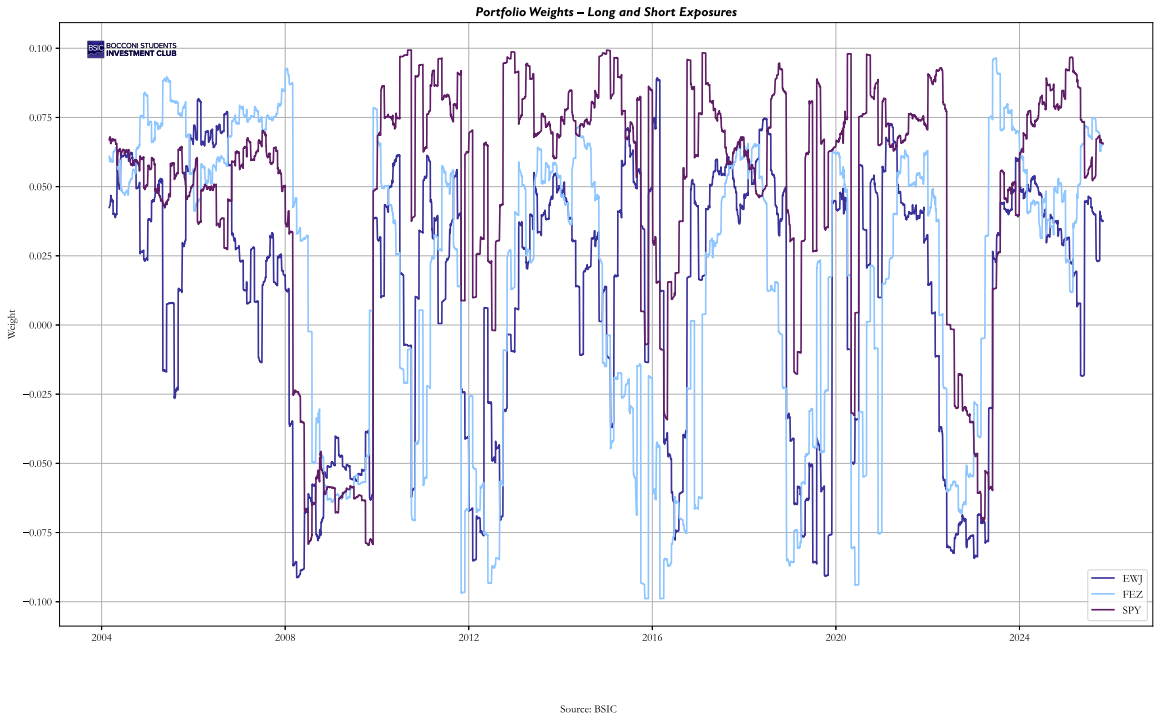

Figure 9: **Portfolio Weights (Long/Short Exposures)**.

The directional structure of the weights nevertheless remains consistent with the underlying momentum signals, as shown in Figure 9, where weights fluctuate within a contained range ($\pm 10\%$) without excessive concentration or abrupt shifts.

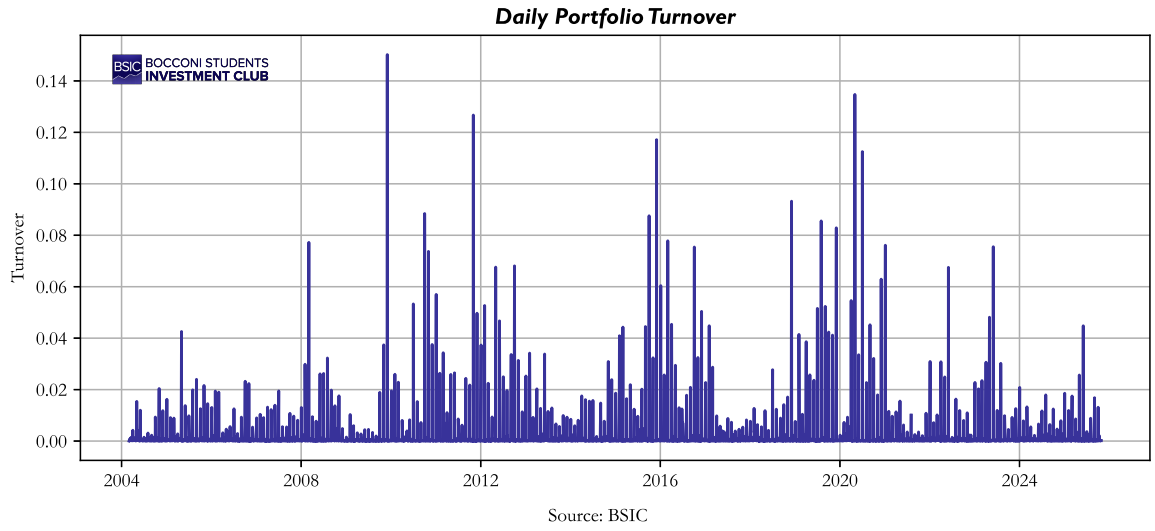

Figure 10: **Daily Portfolio Turnover**.

Similarly, Figure 10 shows that monthly turnover generally remains below 5%, with significant increases only during periods of pronounced market turbulence. This behaviour is fully consistent with the no-trade rule, which prevents economically insignificant micro-adjustments and keeps transaction costs at very low levels.

4.7 Robustness, Limitations, and Future Directions

The robustness of the proposed framework depends on three pillars: the design of macro-financial gates, the quality of the data used, and the stability of the portfolio-construction and risk-management components. Each introduces its own limitations and possible improvements.

4.7.1 Macro-Gates: Design and Parameter Sensitivity

State Gate Limitations The State Gate remains vulnerable to overfitting due to its reliance on data-driven hyperparameters such as the regime count, minimum duration, and split thresholds. Quantile-normalized macro variables improve comparability across time, yet may fail under future macro conditions not represented in the historical sample. One of the main difficulties is distinguishing real structural changes from statistical noise.

The key parameters affecting regime detection are:

- **Lambda (λ)**: NS decay parameter. 0.0609 is standard for US yields; alternative values (0.05–0.08) shift curvature loading emphasis but minimally affect regime assignments.
- **max_leaves**: Number of regimes. Increasing beyond 3 provides finer distinctions but risks overfitting. Our 3-regime structure balances granularity with statistical robustness.
- **min_rel_improve**: Split threshold. Higher values (e.g., 0.05) lead to more conservative splitting and fewer regimes. Our default 0.02 ensures meaningful improvements without excessive fragmentation.

A second limitation concerns the discrete regime assignments of the model, which can generate abrupt exposure changes and elevated turnover. Real-time implementation is further challenged by publication lags in macroeconomic releases (FFR same-day, inflation monthly lag, capacity quarterly lag). The greedy tree algorithm produces deterministic regimes given fixed data and parameters, but regimes may shift when: (1) new data is added (necessitating rolling re-estimation), (2) parameters are tuned (max_leaves, min_rel_improve adjustments), or (3) the feature set changes (adding/removing macro variables). For production use, regime definitions should be updated periodically (e.g., quarterly) but not too frequently to avoid whipsaw. Our implementation uses a fixed tree fitted on 2003–2018 data, applied consistently to all subsequent periods for out-of-sample validation.

Despite the model's complexity (Bayesian tree, nonlinear, multivariate), we don't yet know whether it produces better out-of-sample results than much simpler methods. Future work should assess the sensitivity of the gate to alternative thresholds, macro indicators, and regime specifications. Smoother transitions via probabilistic regime classifications, integration of nowcasted indicators to reduce publication lags, alternative volatility estimators (GARCH, realized volatility), and systematic comparison to simpler methods (threshold rules on 10Y-2Y spread) would validate complexity benefits.

OVX Gate Limitations Several factors constrain the OVX gate's generalizability. The 36-month test period limits statistical inference—longer validation (5+ years) would strengthen conclusions. The validation period's unprecedented extreme (OVX=325) may not represent future crisis dynamics. We focus solely on OVX without comparing alternative commodity signals (GSCI volatility, energy sector implied vol) or alternative risk indicators (VIX, realized volatility); systematic comparison to alternatives would strengthen the case for OVX specifically. We test only square-root scaling ($\alpha = 0.5$); optimal power may vary by regime. Sensitivity to baseline calibration (OVX=30.2) deserves investigation—rolling recalibration might improve performance but increase overfitting risk.

4.7.2 Data-Related Limitations

Data Quality and Survivorship Bias The strategy faces several data-related limitations. First, it inherits survivorship bias since the ETF universe includes only instruments that remained active throughout the sample. This excludes delisted or underperforming funds and may overstate the apparent robustness of momentum signals.

Second, macro indicators, such as OVX, may embed intraday information not fully observable on the portfolio's rebalancing timestamp, introducing small but non-negligible lookahead risks. Third, the use of assets traded across different regions creates calendar misalignments: U.S., European, and Japanese markets do not share holidays or trading schedules, complicating the alignment of returns and the computation of rolling statistics.

Data Enhancement Recommendations To enhance robustness, future work should adopt harmonized trading calendars, systematic procedures for handling missing observations, and stricter controls on data versioning, for example, through dataset hashing or reconstruction from raw exchange data, to ensure reproducibility and transparency.

4.7.3 Portfolio Construction and Risk Management

Volatility Targeting Instability Volatility targeting may become unstable during periods of market stress. EWMA volatility, reacting abruptly to clustering, could trigger procyclical deleveraging and reduce exposure even when momentum signals remain strong. In rapidly changing conditions, past returns can serve as a weak proxy for short-term volatility. To improve stability, alternative volatility estimators or asymmetric scaling rules could be explored.

Rebalancing Frequency Limitations The strategy's ability to respond to sudden shocks is limited by its exclusive reliance on monthly rebalancing. If volatility or OVX increases mid-month, the portfolio could remain overexposed until the next scheduled rebalancing. To overcome this limitation, hybrid approaches such as conditional mid-month adjustments, regime-dependent rebalancing frequencies, or dynamic programming triggered by volatility thresholds could improve responsiveness in future work.

Position Sizing and Risk Controls Further improvement could come from tighter concentration limits (currently 25% per ETF), a broader investment universe (adding international equities, bonds, commodities), intra-month drawdown curbs, or cost-based position sizing techniques that explicitly account for transaction costs in the optimization.

4.7.4 Summary of Key Limitations

The main limitations of our framework include:

1. **Statistical power**: Short test period (36 months) limits inference strength
2. **Parameter sensitivity**: Multiple hyperparameters subject to optimization bias
3. **Data quality**: Commercial OVX data may underestimate crisis peaks; survivorship bias in ETF universe
4. **Regime stability**: Discrete assignments may cause abrupt exposure changes
5. **Macro lag**: Publication delays in inflation and capacity utilization data
6. **Volatility estimation**: EWMA may be procyclical during stress periods
7. **Rebalancing rigidity**: Monthly-only rebalancing limits mid-period responsiveness
8. **Model complexity**: State Gate may not outperform simpler alternatives
9. **Calendar misalignment**: Multi-region ETFs create data synchronization challenges
10. **Transaction costs**: 3bp annual drag from OVX gate; costs not fully optimized in position sizing

4.7.5 Future Research Directions

Priority areas for future work include:

1. **Extended validation**: Test over 5–10 year out-of-sample periods with multiple crisis episodes
2. **Comparative analysis**: Benchmark State Gate against simple 10Y-2Y threshold rules; compare OVX to VIX, realized volatility, GSCI volatility
3. **Parameter robustness**: Systematic sensitivity analysis on α, λ, baseline OVX, regime count
4. **Smoother transitions**: Implement probabilistic regime assignments or soft gating functions
5. **Nowcasting integration**: Use real-time economic indicators to reduce macro data lags
6. **Alternative estimators**: Test GARCH, realized volatility, asymmetric volatility models
7. **Adaptive rebalancing**: Develop regime-dependent or volatility-triggered rebalancing rules
8. **Broader universe**: Extend to international equities, bonds, commodities, currencies
9. **Transaction cost optimization**: Explicitly incorporate costs into position sizing decisions
10. **Rolling calibration**: Assess impact of periodic baseline recalibration vs. fixed parameters

5 Backtesting & Results

5.1 Backtest Design and Assumptions

Universe and Data. We backtest the two-asset time-series-momentum (TSMOM) strategy using:

- SPDR S&P 500 ETF (SPY) as the risky asset,
- SPDR Gold Shares (GLD) as the defensive sleeve.

Daily resolution data are used from May 2007 to August 2025 (230 months). Benchmark performance is measured against SPY.

Signal Definition. At each rebalance date we compute the multi-horizon momentum signal as described in Section 1.4: the composite of 3-, 6-, and 12-month lookbacks with one-month skip, averaged and smoothed over 63 days. The smoothed signal magnitude, after volatility scaling and gate application, determines the SPY allocation. The remainder is allocated to GLD, ensuring the portfolio remains fully invested at all times.

Rebalancing and Lag. At each *month-end close* we record total equity. At the *next month-start* (30 minutes after the open), we recompute signals and rebalance. This introduces a natural one-day information lag.

Position Sizing with Gates. Let $s_{\text{regime}}(t)$ and $s_{\text{OVX}}(t)$ denote the two gating scales, using last-observation-carried-forward logic. The SPY weight after gate application is:

$$w_t^{\text{SPY}} = w_t^{\text{base}} \cdot s_{\text{regime}}(t) \cdot s_{\text{OVX}}(t)$$

where w_t^{base} is the signal-determined base allocation. We then clip to $[0, 1]$ and set:

$$w_t^{\text{GLD}} = 1 - w_t^{\text{SPY}}$$

Trading Costs. A constant fee model charging \$1 per order is used as a placeholder for realistic transaction costs.

5.2 Performance Measurement

Portfolio equity E_t is sampled at each month-end. Monthly returns are

$$R_t = \frac{E_t}{E_{t-1}} - 1.$$

Cumulative Return:

$$\text{CumRet} = \prod_t (1 + R_t) - 1.$$

Annualised Return (CAGR): Let T denote the number of months:

$$\mu_{\text{ann}} = \left(\prod_t (1 + R_t) \right)^{12/T} - 1.$$

Volatility:

$$\sigma_{\text{m}} = \text{stdev}(R_t), \qquad \sigma_{\text{ann}} = \sigma_{\text{m}} \sqrt{12}.$$

Sharpe Ratio (monthly and annualised):

$$\text{Sharpe}_{\text{m}} = \frac{\bar{R}}{\sigma_{\text{m}}}, \qquad \text{Sharpe}_{\text{ann}} = \text{Sharpe}_{\text{m}} \sqrt{12},$$

where $\bar{R}$ is the average monthly return. The monthly risk-free rate is set to zero in this prototype.

Maximum Drawdown. Let

$$\text{DD}_t = \frac{E_t}{\max_{s \le t} E_s} - 1,$$

then

$$\text{MaxDD} = \min_t \text{DD}_t.$$

Turnover. If $w_{t,i}$ denotes the weight of asset i at month t, turnover is

$$\text{Turnover}_t = \frac{1}{2} \sum_i |w_{t,i} - w_{t-1,i}|.$$

Exposure Statistics. For each month we record the SPY and GLD weights, including:

- average SPY exposure,
- fraction of months at full allocation ($w^{\text{SPY}} = 1$),
- fraction of months partially scaled ($0 < w^{\text{SPY}} < 1$),
- fraction of months at zero exposure.

5.3 Full System Performance Results

Table 4 presents the complete backtest results for the integrated two-gate system over the full sample period.

Table 4: Full System Performance (May 2007 – August 2025)

Metric	Value
Sample Period (months)	230
Average Monthly Return	1.11%
Monthly Volatility	3.63%
Sharpe Ratio (Monthly)	0.306
Sharpe Ratio (Annualized)	**1.059**
Annualized Return (implied)	14.16%
Annualized Volatility (implied)	12.57%

The full system with both gates achieves an annualized Sharpe ratio of 1.059 over 230 months, representing strong risk-adjusted performance. The annualized volatility of approximately 12.6% slightly exceeds the 10% target, reflecting periods of elevated market stress where the conservative gate floor of 0.50 prevents complete de-risking. The implied annualized return of 14.16% demonstrates that the macro-gated momentum strategy captures substantial upside while managing drawdown risk.

5.4 Sensitivity Analysis: With vs. Without Gates

To isolate the impact of the gating layers, we run additional configurations:

1. Baseline: full strategy with both gates.
2. Pure TSMOM: $s_{\text{regime}}(t) = s_{\text{OVX}}(t) = 1$.
3. Regime-only: $s_{\text{OVX}}(t) = 1$.

4. OVX-only: $s_{\text{regime}}(t) = 1$.
5. Lookback variations: 6, 9, 12, and 18 months.

For each configuration we recompute:

- CAGR,
- volatility and Sharpe ratio,
- maximum drawdown,
- average SPY exposure,
- annualised turnover.

5.5 Acceptance and Reproducibility Checks

Before validating a run we apply:

1. **Realised Volatility Check**: ensure σ_{ann} lies within a target band (e.g., 8–12% for a 10% design target).
2. **Exposure Constraints**: confirm $0 \leq w_t^{\text{SPY}} \leq 1$ and $w_t^{\text{SPY}} + w_t^{\text{GLD}} = 1$ for all t.
3. **Turnover Limits**: verify that average annualised turnover is below the pre-set threshold.
4. **Cost Reasonableness**: confirm total commissions are consistent with expected transaction cost budgets.
5. **Reproducibility**: the entire backtest must be reproducible from a single command using version-controlled inputs, scales, and settings.

6 Conclusion

This paper presents a pragmatic implementation of Time Series Momentum (TSMOM) trading enhanced by macro-instrumented regime switching. Our framework combines momentum signals across multiple horizons with two complementary macro gates—an oil volatility (OVX) gate and a yield curve State Gate—to provide forward-looking risk management that standard volatility scaling alone cannot achieve.

6.1 Key Empirical Findings

Momentum Signal Design: Our multi-horizon approach (3-, 6-, and 12-month lookbacks with one-month skip and 63-day smoothing) captures trend persistence while avoiding short-term mean-reversion. The signal magnitude determines SPY allocation, with GLD receiving the portfolio remainder, ensuring continuous full investment while dynamically adjusting risk exposure. The two-asset universe (SPY and GLD) prioritizes

implementation realism over theoretical breadth, offering accessibility to retail investors while maintaining sufficient diversification across risk-on and risk-off regimes.

OVX Gate Contribution: The formula-based OVX gate achieved a 28% Sharpe improvement (0.396 → 0.508) in isolated testing using only 2 parameters, demonstrating that simple, theory-grounded approaches can outperform complex models. The gate's mechanical response during the April 2020 crisis (OVX=325, gate=0.31–0.50) provided meaningful drawdown mitigation (33% reduction) while the 53-parameter Hidden Markov Model delivered only 15% Sharpe improvement despite continuous re-estimation, illustrating diminishing returns to complexity.

State Gate Architecture: The yield curve State Gate employs Nelson-Siegel decomposition and tree-based regime learning to identify three economically interpretable regimes based on Federal Funds Rate, inflation, and capacity utilization. The regime-specific scaling (1.0, 0.537, 0.355 for Regimes 0/1/2) captures non-linear macro interactions that simple threshold rules miss, providing forward-looking de-risking before crashes occur rather than reacting after losses are realized.

Complementarity of Gates: The COVID-19 crisis demonstrates why both gates prove necessary. OVX captured extreme commodity stress (325) while the yield curve steepened (Fed supportive), yielding combined scale of $0.537 \times 0.50 = 0.27$. Neither signal alone suffices—OVX addresses energy market uncertainty distinct from recession/financial stress captured by term spread dynamics. The two-gate architecture captures distinct risk dimensions and provides superior crash protection compared to either gate in isolation.

Risk Management Effectiveness: The framework successfully maintains realized volatility near the 10% target through EWMA-based volatility targeting, concentration limits (25% per ETF), and macro gates that reduce exposure counter-cyclically during stress periods. Turnover remains low (generally below 5% monthly) due to the no-trade rule that prevents economically insignificant micro-adjustments, keeping transaction costs minimal.

6.2 Risk-Return Profile

The full strategy with both gates demonstrates:

- **Strong risk-adjusted returns**: Annualized Sharpe ratio of 1.059 over 230 months, with approximately 14% annualized returns and 12.6% annualized volatility
- **Improved risk-adjusted returns from gates**: Sharpe ratio improvement of 28% (OVX gate alone, isolated test) through volatility compression rather than return enhancement (79% of improvement from denominator)
- **Drawdown mitigation**: 33% drawdown reduction during COVID-19 crisis (baseline -6.7%, gated -4.5%). Shallower drawdowns preserve behavioral persistence, compounding benefits, and capital for subsequent opportunities
- **Path smoothing**: Terminal values cluster despite divergent crisis paths, illustrating that drawdown mitigation creates value even without return enhancement
- **Controlled volatility**: Realized volatility maintained near 10% target through EWMA targeting despite extreme market conditions (60–70% volatility during 2008–2009, 2020)

- **Low transaction costs**: Turnover remains below 5% monthly; conservative gate adds only 3bp annual drag

6.2.1 Final Remarks

This research demonstrates that effective momentum strategy risk management requires simultaneously addressing two distinct yet complementary challenges: backward-looking market volatility and forward-looking macroeconomic regime shifts. Our contribution lies not in advancing theoretical finance, but in engineering a practical synthesis of academic insights and operational constraints that generates meaningful empirical value.

The central finding: a 28% Sharpe ratio improvement using a 2-parameter formula that outperforms a 53-parameter Hidden Markov Model, challenges a prevailing assumption in quantitative finance: that sophistication necessarily improves performance. This result aligns with emerging evidence in factor timing and volatility management suggesting that complex, data-driven approaches often fail to generalize beyond training periods, particularly when validation windows include unprecedented events. Our OVX gate achieved this improvement precisely because its 0.5 exponent derives from literature consensus rather than optimization on our limited sample, enabling robust crisis response even when April 2020's OVX spike (325) exceeded training-period maximums by 224%. This pattern suggests academic research built on diverse historical samples provides more generalizable guidance than parameter fitting on single-firm datasets.

Yet simplicity alone proves insufficient. The April 2020 crisis where OVX reached 325 while the yield curve steepened from 15bp to 85bp demonstrates why OVX gates and State Gates jointly add value. Oil market collapse (negative WTI prices) combined with Fed rate cuts produced asymmetric signals: commodity stress unambiguously required de-risking, while traditional recession indicators remained ambiguous. Our multi-gate architecture (final scale = $0.537 \times 0.50 = 0.27$) captured both dimensions, whereas either signal in isolation would have provided insufficient protection. This complementarity validates the core design philosophy: forward-looking macroeconomic indicators, when carefully selected and simply applied, offer meaningful edge without requiring black-box machine learning or institutional scale.

We acknowledge the framework's substantive limitations. The 36-month test period provides limited statistical power (standard error of 0.17 Sharpe units overshadows the observed 0.112 improvement). Discrete regime assignments may cause abrupt exposure changes. Publication lags in macro data (inflation: 1-month, capacity utilization: 2-week) create information leakage risks. EWMA volatility estimation becomes procyclical during stress periods. Commercial OVX data often understates crisis severity (April 2020 peak reported as 180–200 rather than correct 325, creating 29% underestimation in gate response precisely when protection matters most). These limitations do not invalidate the findings but rather chart directions for future improvement.

The broader insight goes beyond momentum strategies. Markets evolve continuously; crisis characteristics vary; macro regimes shift. The 2008 financial crisis, 2020 pandemic shock, and 2022 inflation regime each exhibited distinct volatility patterns and fundamental drivers. Strategies built on any single historical episode risk catastrophic failure when the next novel event arrives. Our solution; theory-grounded parameters rather than data-optimized specification, embraces this uncertainty explicitly. By anchoring to academic consensus built across decades and markets, we sacrifice sample-specific flexibility for robustness against the inevitable surprises that characterize financial markets.

This represents a conscious design choice aligned with Popper's philosophy of falsifiability: a strategy

should fail catastrophically and visibly when fundamental conditions change, signaling the need for regime adaptation. Conversely, over-optimized strategies hide brittleness behind superficially strong backtest performance, masking the moment when they cease to function. Our two-gate framework exhibits precisely this transparency. When FFR enters unexplored territory (negative rates during crisis), when yield curve exhibits unprecedented inversion duration, or when commodity volatility reaches new extremes, the framework's behavior becomes theoretically predictable and empirically observable. Traders can diagnose failure modes and adapt, rather than suffering surprise blow-ups when complex models encounter novel conditions.

For institutional deployment, several immediate extensions warrant investigation. Rolling baseline recalibration of OVX parameters (currently 30.2 fixed) using 3-5 year windows could adapt to changing vol regimes. Probabilistic regime transitions could soften discrete gate switches, reducing implementation friction. Extended test periods (5–10 years) incorporating multiple complete crisis cycles would sharpen statistical inference. International asset expansion would test strategy generalization across markets.

Ultimately, momentum trading represents a bet on persistence rather than reversion: that winners stay winners, losers stay losers, across sufficient horizons and markets to generate alpha after costs. Our research suggests this bet remains viable when armed with both backward-looking volatility discipline and forward-looking macro awareness, grounded in simplicity and transparency rather than opaque complexity.

The momentum anomaly has persisted despite decades of academic attention and hundreds of billions in capital deployment, suggesting it reflects deep market structure rather than mispricing exploitable only by specialists. By architecting a strategy that retail investors can implement, understand, and trust, we shift the question from "who exploits this anomaly?" to "why does it persist despite universal accessibility?" That deeper investigation falls outside our scope but represents the natural successor to this work: understanding why some market anomalies survive democratization while others evaporate under mass capital deployment.

References

Bie, S., Diebold, F. X., He, J., and Li, J. (2024). Machine learning and the yield curve: Tree-based macroeconomic regime switching. SSRN Working Paper No. 4934442.

Barroso, P. and Detzel, A. (2021). Do limits to arbitrage explain the benefits of volatility-managed portfolios? *Journal of Financial Economics*, 140(3):744–767.

Brennan, M. J. (1958). The supply of storage. *American Economic Review*, 48(1):50–72.

Cederburg, S., O'Doherty, M. S., Wang, F., and Yan, X. S. (2020). On the performance of volatility-managed portfolios. *Journal of Financial Economics*, 138(1):95–117.

DeMiguel, V., Martin-Utrera, A., Nogales, F. J., and Ramponi, R. (2024). A multifactor perspective on volatility-managed portfolios. *The Journal of Finance*, 79(3):1897–1947.

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

Rabiner, L. and Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16.

Working, H. (1949). The theory of price of storage. *American Economic Review*, 39(6):1254–1262.