

Leverage Direction Matters: Cross-Asset Evidence on GARCH Model Selection and Volatility Targeting*

Yi-Hao Lai[†]

May 2026 (v3.3)

Abstract

The direction of the asymmetric volatility response—measured by the GJR-GARCH γ parameter—carries distinct economic content that varies systematically across asset classes. Using daily data for seven primary assets (equities, gold, bonds, emerging markets, cryptocurrency) over 2017–2025 (with 2026 reserved for out-of-sample validation), extended to 26 assets in validation analyses, we establish a single central proposition: the sign of γ identifies the dominant price-driving mechanism of an asset—balance-sheet leverage for equities ($\gamma > 0$), fear-driven flight-to-quality for gold ($\gamma < 0$ conditional on stress regimes), and coupon-dominated pricing for bonds ($\gamma \approx 0$). Gold’s inverted leverage is *regime-dependent* rather than unconditional: the canonical rolling-window mean is statistically indistinguishable from zero (mean $\gamma = +0.002$, HAC $t = +0.15$), while a regime decomposition concentrates inversion in fear-driven rallies and reverts to standard leverage during liquidation-driven declines ($t = -3.79$, $p < 0.001$). This economic content has two empirical manifestations that we test rather than assume. First, γ -based model selection delivers out-of-sample forecasting gains only when asymmetry is statistically significant ($t > 1.65$ on the quarterly-mean HAC statistic), yielding 6/6 correct classifications on the disjoint 2024–2025 sample using 2023 γ estimates

*All errors are my own. Data and replication code are available upon request.

[†]Department of Finance, Da-Yeh University. Corresponding author. Email: yhlai@mail.dyu.edu.tw

(small cross-section acknowledged, $N = 6$). Second, within homogeneous equity markets, γ predicts whether volatility targeting (VT) acts as trend-following or contrarian (Spearman $\rho = 0.886$, $p = 0.019$, $N = 6$); this mapping is a genuine domain restriction that does not survive extension to heterogeneous asset classes ($\rho = -0.448$, $p = 0.14$, $N = 12$). VT’s drawdown-reduction benefit is meanwhile universal and driven by base volatility level ($\rho = 0.944$, $N = 5$; $\rho = 0.83$, $N = 14$), independent of γ . Together these results position leverage direction as an economically interpretable asset-classification device—not a general-purpose forecasting rule.

Keywords: GARCH, leverage effect, gold, asymmetric volatility, Value-at-Risk, Basel III, volatility targeting, cross-asset, regime-dependent

JEL Classification: C22, C53, G01, G11, G17

1 Introduction

The GARCH family of models (Bollerslev, 1986) remains the workhorse for daily volatility forecasting in financial markets. A substantial literature has developed around asymmetric extensions—most notably the GJR-GARCH of Glosten et al. (1993) and the EGARCH of Nelson (1991)—designed to capture the empirical observation that negative equity returns tend to increase future volatility more than positive returns of equal magnitude. This “leverage effect,” first noted by Black (1976) and formalized by Christie (1982), is well-documented for equity markets and has become a standard consideration in volatility model specification.

However, the implicit assumption in much of the applied GARCH literature—that asymmetric models uniformly improve upon their symmetric counterparts—warrants closer examination when moving beyond equities. While Chevallier and Ielpo (2017) document inverted asymmetric volatility in gold and several agricultural commodities, the implications of this reversal for model selection, risk management, and portfolio construction have not been systematically explored.

This paper advances *one central contribution*: the sign of γ carries distinct economic

content that identifies an asset’s dominant price-driving mechanism, not merely a technical parameter for GARCH specification. Balance-sheet leverage ($\gamma > 0$ for equities), fear-driven flight-to-quality conditional on stress regimes ($\gamma < 0$ for gold), and coupon-dominated pricing ($\gamma \approx 0$ for bonds) are three qualitatively different mechanisms by which shocks transmit into conditional variance. Reading γ as an economic signature—rather than as a nuisance parameter—is what the cross-asset evidence in this paper supports.

The central proposition is developed on seven primary assets spanning equities (SPY, QQQ, EEM), precious metals (GLD, SLV), government bonds (TLT), and cryptocurrency (BTC-USD), and validated on up to 26 assets. Gold’s inverted asymmetry warrants particular care: it is *regime-dependent*, not unconditional. The canonical full-replication estimate of the rolling-window mean is statistically indistinguishable from zero (mean $\gamma = +0.002$, HAC $t = +0.15$, 67% of quarterly windows negative), whereas a regime decomposition concentrates inversion in fear-driven rallies and reverts to standard leverage during liquidation-driven declines ($t = -3.79$, $p < 0.001$ for the regime difference; Section 5.3). We therefore treat the inverted-leverage channel as an economic phenomenon conditional on regime—not a permanent property of the asset class—and argue explicitly against reading unconditional γ estimates for gold as evidence of a stable inversion.

This economic content has two empirical manifestations that we test rather than assume. The first concerns forecasting: γ -based model selection delivers genuine out-of-sample gains only when asymmetry is statistically significant. Under the t -statistic rule ($t > 1.65$ on the quarterly-mean HAC statistic), the prescribed model is never significantly beaten in any of the eleven Diebold-Mariano comparisons in the primary in-sample panel (Table 3), and it delivers 6/6 correct classifications on the disjoint 2024–2025 sample using 2023 γ estimates; BTC-USD illustrates that a statistically significant γ ($t = +2.88$) does not guarantee an OOS forecasting edge—its two models are statistically indistinguishable out of sample. We flag the small cross-section ($N = 6$) explicitly; we do not claim a general-purpose forecasting rule.

The second manifestation concerns portfolio construction. Within homogeneous equity markets, γ predicts whether volatility targeting (VT) acts as trend-following or

contrarian (Spearman $\rho = 0.886$, $p = 0.019$, $N = 6$; out-of-sample $\rho = 0.821$ using γ from 2010–2017 to predict trend beta in 2018–2026). This mapping is a genuine domain restriction: it does not extend to heterogeneous asset classes ($\rho = -0.448$, $p = 0.14$, $N = 12$; Section 5.3 states the domain conditions). VT’s drawdown-reduction benefit is meanwhile universal across all tested assets and driven by base volatility level (Pearson $\rho = 0.944$, $p = 0.016$, $N = 5$ primary; $\rho = 0.83$, $p = 0.0002$, $N = 14$ extended), independent of γ . Reading these two results jointly, leverage direction is best understood as an economically interpretable classification device: it delineates *when* asymmetric structure has portfolio-level economic value—not as a rule of thumb to be mechanically applied.

Supplementary evidence on cross-market transmission and auxiliary robustness checks is reported in the online supplement.

The remainder of the paper is organized as follows. Section 2 reviews the literature. Section 3 describes data and methodology. Section 4 presents empirical results. Section 5 discusses implications and limitations. Section 6 concludes.

2 Literature Review

The paper sits at the intersection of three literatures. The first studies asymmetric volatility models: GARCH captures volatility clustering, while GJR-GARCH and EGARCH allow negative and positive shocks to affect variance differently (Bollerslev, 1986; Glosten et al., 1993; Nelson, 1991). In equities, this leverage effect is standard and economically tied to balance-sheet amplification (Black, 1976; Christie, 1982; Engle and Siriwardane, 2018). A smaller literature shows that non-equity assets can reverse the sign of the asymmetry, especially for gold and some commodities, but stops short of turning that sign information into a model-selection rule (Chevallier and Ielpo, 2017; Batten et al., 2010; Chang et al., 2021).

The second literature concerns risk management. Basel backtesting makes the distributional assumption as important as the variance equation itself (BCBS, 2006, 2019; Kupiec, 1995; Christoffersen, 1998). Student- t and skewed- t innovations often improve

tail coverage relative to Gaussian specifications (Bollerslev, 1987; Hansen, 1994; Kuester et al., 2006). The third literature studies volatility targeting: inverse-volatility scaling improves drawdown control and sometimes improves Sharpe ratios, though the evidence is criterion- and asset-dependent (Fleming et al., 2001, 2003; Moreira and Muir, 2017; Harvey et al., 2018; Cederburg et al., 2020; Bozovic, 2024; DeMiguel et al., 2024; Hood and Raughtigan, 2025; Xu, 2024). Our contribution is to connect these strands through a single empirical object, the sign and magnitude of γ , and then test where that information matters for forecasting, VaR, and portfolio construction.

3 Data and Methodology

3.1 Data

We use daily adjusted closing prices for seven exchange-traded assets: SPDR S&P 500 ETF Trust (SPY), Invesco QQQ Trust (QQQ), iShares MSCI Emerging Markets ETF (EEM), SPDR Gold Shares (GLD), iShares Silver Trust (SLV), iShares 20+ Year Treasury Bond ETF (TLT), and Bitcoin (BTC-USD). Data are sourced from Yahoo Finance adjusted close prices for the period January 2017 through March 2026.

Data source justification. Yahoo Finance provides the daily history for all assets. This choice is standard in empirical asset-pricing practice and is defensible here because the sample consists of large, liquid ETFs and Bitcoin, not microcaps where stale prices and delayed corporate-action adjustments are more problematic (Bali et al., 2016). We validate the feed by spot-checking crisis dates against Bloomberg, matching Yahoo VIX to official series, and confirming that return moments line up with overlapping benchmark studies (Moreira and Muir, 2017; Cederburg et al., 2020). Daily returns are computed as simple percentage changes. We nevertheless treat institutional-grade replication as the relevant production standard.

Table 1 reports descriptive statistics. All return series reject normality (Jarque-Bera $p < 0.001$), exhibit significant ARCH effects (Engle’s LM test (Engle, 1982), $p < 0.001$),

and are stationary (ADF $p < 0.001$). Excess kurtosis ranges from 3.52 (GLD) to 14.61 (SPY), motivating the consideration of fat-tailed distributions for VaR estimation.

Sample period justification. The primary sample spans January 2017 to March 2026, partitioned into three non-overlapping windows: in-sample training (2017-01 to 2022-12), main out-of-sample (2023-01 to 2024-12), and out-of-sample validation (2025-01 to 2026-03). This is shorter than some long-horizon GARCH studies, but each forecast is estimated on a rolling two-year window, and robustness checks extend the effective estimation history with larger windows back to 2006. The key object is not long-horizon return predictability but the sign and usefulness of γ , which is stable across the window sizes we test and remains qualitatively unchanged in stressed subperiods such as 2020–2022 and the 2025–2026 Iran/Hormuz episode.

3.2 Volatility Models

We compare the symmetric GARCH(1,1) of [Bollerslev \(1986\)](#), $\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$, against the GJR-GARCH(1,1) of [Glosten et al. \(1993\)](#):

$$\sigma_t^2 = \omega + (\alpha + \gamma\mathbf{1}_{\varepsilon_{t-1} < 0})\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \quad (1)$$

where γ captures asymmetric responses: $\gamma > 0$ implies standard leverage (negative returns increase variance), $\gamma < 0$ implies inverted leverage.

Mean specification. We adopt the zero-mean specification $r_t = \sigma_t\varepsilon_t$ throughout. This choice is grounded in the well-documented finding that the conditional mean of daily equity returns is negligible relative to conditional variance: for a typical daily mean of 0.04% and daily standard deviation of 1%, the mean-to-volatility ratio is approximately 0.04, making its contribution to variance dynamics immaterial ([Bollerslev, 1986](#); [Hansen and Lunde, 2005](#)). [Francq and Zakoian \(2004\)](#) show that quasi-maximum likelihood estimation of GARCH parameters remains consistent regardless of mean specification when the true conditional mean is small relative to conditional variance. As a robustness

check, we re-estimate all models with an AR(1) mean specification $r_t = \mu + \phi r_{t-1} + \sigma_t \varepsilon_t$; results are qualitatively unchanged—QLIKE rankings, gamma sign classifications, and DM test conclusions are all preserved (Section 4.6, Model Specification).

All models are estimated with Normal quasi-likelihood via the `arch` package (Shepard, 2023). For VaR, we additionally consider Student- t innovations with fixed degrees of freedom.

3.3 Rolling Estimation

We employ a rolling window approach with re-estimation at each forecast origin. The primary window size is $w = 504$ trading days (approximately 2 calendar years), which satisfies the minimum sample size recommendation for GARCH estimation (≥ 500 observations; see Hwang and Valls Pereira 2006) while avoiding contamination from distant regime changes. Robustness checks with $w \in \{252, 1000, 2000, 3000, 5000\}$ (reported in the online supplement) confirm that QLIKE rankings are qualitatively robust to window choice, though $w = 504$ produces the best QLIKE during high-volatility periods (2020–2021) while larger windows ($w \geq 2000$) perform marginally better during calm periods (2023–2024). The gamma sign classification is invariant to window size across all tested values.

For each out-of-sample date t , we estimate the model using data from $t - w$ to $t - 1$, and produce a one-step-ahead variance forecast $\hat{\sigma}_t^2$. The realized variance proxy is the squared daily return r_t^2 .

3.4 Evaluation Criteria

3.4.1 Statistical Loss Functions

The primary evaluation metric is the QLIKE loss function (Patton, 2011), defined as the Gaussian quasi-log-likelihood contribution:¹

$$\text{QLIKE} = \frac{1}{T} \sum_{t=1}^T \left(\frac{r_t^2}{\hat{\sigma}_t^2} + \ln \hat{\sigma}_t^2 \right) \quad (2)$$

QLIKE is robust to noise in the realized variance proxy and ranks forecasts consistently regardless of the proxy used, provided the proxy has equal bias across models.

3.4.2 Diebold-Mariano Test

We test the null hypothesis of equal predictive accuracy between models using the Diebold-Mariano test (Diebold and Mariano, 1995) with Newey-West HAC standard errors (5 lags):

$$DM = \frac{\bar{d}}{SE_{NW}(\bar{d})} \sim N(0, 1) \quad (3)$$

where $d_t = L(r_t^2, \hat{\sigma}_{1,t}^2) - L(r_t^2, \hat{\sigma}_{2,t}^2)$ is the loss differential using QLIKE.

3.4.3 VaR Backtesting

Value-at-Risk at confidence level α is computed as:

- **Normal:** $\text{VaR}_\alpha = \Phi^{-1}(\alpha) \cdot \hat{\sigma}_t$
- **Student-t:** $\text{VaR}_\alpha = t_\nu^{-1}(\alpha) \cdot \sqrt{(\nu - 2)/\nu} \cdot \hat{\sigma}_t$

where ν is the degrees of freedom (fixed at 5 in our baseline, consistent with typical empirical estimates for financial returns). The choice of $\text{df} = 5$ is validated out-of-sample:

Section C demonstrates that Basel III Green Zone compliance is achieved for all $\text{df} \leq$

¹This formulation is equivalent to $-\ln f(r_t | \hat{\sigma}_t^2)$ under Gaussian quasi-likelihood, up to an additive constant. Values are negative for daily returns because $\ln \hat{\sigma}_t^2 \approx \ln(10^{-4}) \approx -9.2$. **Lower (more negative) values indicate better forecasts.** An alternative convention in the literature defines the Patton (2011) centered loss as $L^*(\sigma^2, h) = \sigma^2/h - \ln(\sigma^2/h) - 1$, which produces values centered near 1.0. Both formulations preserve forecast rankings; we adopt the quasi-log-likelihood form throughout the main text and the online supplement for consistency with maximum likelihood estimation.

6, and that jointly estimating df produces *worse* performance by over-adapting to quiet markets. The fixed df = 5 is thus a conservative choice that is robust across the empirical range $df \in [4, 7]$.

We apply Kupiec’s (1995) unconditional coverage test (LR statistic, χ^2 with 1 d.f.) and Christoffersen’s (1998) independence test to assess violation clustering. Annual violation counts are classified per the Basel III traffic light system (Green: 0–4, Yellow: 5–9, Red: ≥ 10 violations per approximately 250 trading days).

3.4.4 Leverage Direction Analysis

To analyze the temporal stability of the leverage direction, we estimate GJR-GARCH on semi-overlapping quarterly windows (504-day windows advanced by 63 trading days) and collect the gamma estimates. We test:

- **H0:** $E[\gamma] \geq 0$ vs. **H1:** $E[\gamma] < 0$ (inverted leverage)

using a one-sided t-test on the quarterly gamma series.

3.5 Volatility Targeting

Following Moreira and Muir (2017), the volatility-managed portfolio weight is:

$$w_t = \frac{\sigma_{\text{target}}}{\hat{\sigma}_t} \quad (4)$$

We set $\sigma_{\text{target}} = 10\%$ annualized for the GARCH-based analysis, and $\sigma_{\text{target}} = 12\%$ for the VIX-based analysis (Section 5.5), reflecting the VIX’s systematic upward bias from the embedded variance risk premium. Weights are smoothed with a 5-day moving average and clipped to $[0, 1.5]$. Post-hoc transaction cost analysis confirms that the estimated annual turnover of 756% generates cost drag of only 8–15 basis points at typical SPY bid-ask spreads.

4 Empirical Results

4.1 Data Characteristics

Table 1 summarizes the distributional properties of daily returns for our seven assets over the full sample period (January 2017 to December 2025); as described in Section 3.1, this sample is partitioned into an in-sample training window (2017-01 to 2022-12), a main OOS window (2023-01 to 2024-12), and an OOS validation window (2025-01 to 2026-03). All return series exhibit significant excess kurtosis, ranging from 3.52 (GLD) to 14.61 (SPY), and decisively reject normality (Jarque-Bera $p < 0.001$). The ARCH LM test confirms significant conditional heteroskedasticity in all series ($p < 0.001$), justifying GARCH-family models.

Return skewness varies across assets: SPY (-0.315), QQQ (-0.189), GLD (-0.296), and EEM (-0.584) show negative skewness, while TLT ($+0.155$) is approximately symmetric. A naive model selection approach based on skewness would suggest asymmetric GARCH for all negatively skewed assets. As we demonstrate in Section 4.2, this criterion is misleading for gold: despite negative skewness, gold’s conditional variance responds asymmetrically to *positive* shocks.

Annualized volatility spans an order of magnitude, from 14.6% (GLD) to 57.3% (BTC-USD), motivating our analysis of VT effectiveness as a function of base volatility (Section 4.5). The extreme minimum returns—SPY at -10.94% (March 2020 COVID crash) and BTC-USD at -37.17% —underscore the importance of fat-tailed distributions for VaR estimation (Section C).

4.2 Leverage Direction Across Asset Classes

4.2.1 Rolling Gamma Estimates

We estimate the asymmetry parameter γ of GJR-GARCH(1,1) using a rolling window of 504 trading days, updating quarterly. Table 2 reports cross-sectional statistics for each asset. Figure 1 illustrates the time-series evolution of these rolling estimates across assets.

The most striking finding is the **consistent sign difference across asset classes**.

For equities (SPY, QQQ), γ is uniformly positive across all quarterly estimates, indicating the well-documented leverage effect. Gold (GLD) is the most subtle case: the canonical replication ($n = 58$ quarterly rolling windows, $w = 504$, 2010–2026, HAC 8 lags) yields a mean rolling γ of $+0.002$ (HAC $t = +0.15$), statistically indistinguishable from zero, with 67% of quarterly windows negative.² The unconditional mean therefore does *not* establish inverted leverage for gold; rather, the inversion is **regime-dependent**—concentrated in fear-driven rally periods and reversing to standard leverage in liquidation-driven declines ($t = -3.79$, $p < 0.001$ for the regime difference; Section 5.3). Conditional on regime, the inverted-leverage channel—where *positive* returns increase conditional variance—has not been documented in the prior GARCH literature as a systematic phenomenon.

Treasury bonds (TLT) show near-zero γ (mean = -0.005 , std = 0.044 , 69% of windows negative), indicating no significant asymmetry. Bitcoin (BTC-USD) displays mild but statistically significant standard leverage (mean $\gamma = +0.072$, HAC $t = +2.88$), though with higher cross-window instability (std = 0.105 , 25% negative windows) and occasional sign reversals.

The inverted leverage in gold is consistent with gold’s role as a safe-haven asset (Baur and McDermott, 2010; Baur and Lucey, 2010). During market stress, fear-driven buying drives prices upward, creating heightened uncertainty that elevates conditional variance. Crucially, this mechanism operates opposite to equity leverage, where falling prices increase debt-to-equity ratios and firm risk (Black, 1976; Christie, 1982). The absence of asymmetry in bonds is economically intuitive: interest rate changes do not systematically alter sovereign credit risk in either direction.

4.2.2 Implications for Model Selection

Table 3 presents GARCH(1,1) versus GJR-GARCH(1,1) QLIKE comparisons across all asset-period combinations. The results reveal a clear pattern: GJR-GARCH significantly

²An earlier draft reported mean $\gamma = -0.067$ (HAC $t = -5.79$, 93% negative windows) for the same specification; the canonical re-estimation (documented in the replication package) reverses the sign of the mean to $+0.002$ and renders it insignificant. We adopt the re-estimated values throughout and treat the unconditional rolling-mean evidence as inconclusive; the economically interpretable result is the regime decomposition below.

outperforms only when γ is consistently positive and above approximately 0.10.

For SPY ($\gamma \approx +0.13$), GJR achieves significantly lower QLIKE (DM $p = 0.003$ for 2023–2024, $p = 0.048$ for 2025). For QQQ ($\gamma \approx +0.12$), GJR’s advantage is directionally consistent but not significant ($p = 0.181$ for 2023–2024, $p = 0.086$ for 2025). For GLD—whose mean rolling γ is statistically zero—symmetric GARCH *significantly outperforms* GJR in 2023–2024 ($\Delta = +0.39\%$, $p = 0.001$), with the same direction in 2025 ($p = 0.070$): imposing an asymmetry term on an asset without stable asymmetry actively hurts forecasts. For TLT ($\gamma \approx 0$), EEM, and BTC-USD, Diebold-Mariano tests fail to reject equal accuracy at the 5% level.

This finding has practical implications: the conventional approach of selecting asymmetric GARCH models based on return skewness can be misleading. GLD exhibits substantial negative skewness (-0.296) yet has inverted leverage—skewness-based selection would incorrectly prescribe GJR. We propose that **the sign and magnitude of estimated γ** should replace skewness as the primary criterion.

4.2.3 Robustness

The γ sign classification is robust to window size (results qualitatively identical with $w = 252$ and $w = 756$), estimation period (2007–2026 for SPY shows the same pattern), and model specification (EGARCH leverage parameter shows consistent sign alignment with GJR γ).

Because our quarterly gamma estimates use overlapping windows (504-day windows stepped by 63 days, yielding 87.5% overlap between consecutive estimates), serial dependence in the gamma series is a first-order concern. As [Harri and Brorsen \(2009\)](#) emphasize, overlapping observations inflate the effective sample size relative to the number of independent data points, biasing naive t -statistics upward. We address this in two ways.

First, we employ Newey-West HAC standard errors ([Newey and West, 1987](#)) with 8 lags (chosen as $\lceil 0.75 \times T^{1/3} \rceil$ where T is the number of quarterly gamma estimates), which accounts for the moving-average structure induced by the overlap. All t -statistics

in Table 2 are HAC-corrected; for SPY, the corrected statistic remains strongly significant ($t = +11.08$), and for GLD the corrected statistic ($t = +0.15$) confirms that the unconditional mean carries no significance to begin with.

Second, and most importantly, we repeat the cross-sectional analysis using strictly non-overlapping 5-year windows (2006–2010, 2011–2015, 2016–2020, 2021–2025). While the smaller number of cross-sections ($N = 4$) limits formal inference, the gamma rankings for the unambiguous classes are qualitatively preserved: SPY exhibits positive mean γ in all four windows (range: +0.14 to +0.27) and TLT remains near zero in all windows ($|\bar{\gamma}| < 0.02$). Gold’s non-overlapping window means were negative in an earlier-draft vintage (range -0.092 to -0.031); given the sign reversal of the quarterly-mean estimate, we treat gold’s unconditional sign as unresolved at the window level and rest the gold claim entirely on the regime decomposition of Section 5.3.

We conclude that while the 87.5% overlap in our primary analysis inflates the number of gamma observations, it does not drive the qualitative findings for equities and bonds. The overlap provides finer temporal resolution for regime analysis (Section 5.3).

Within the canonical 2010–2026 window, GLD’s γ is negative in 67% of quarterly estimates. An extended OOS validation (January 2025–March 2026, $N = 300$) confirms the equity-side findings: the gamma taxonomy is consistent across the seven assets, GJR outperforms for SPY with widening improvement ($DM = 4.82$, $p < 0.0001$), and 12/VIX reduces MDD by 43% during the April 2025 selloff ($VIX = 52.3$).

The QLIKE ranking is robust to volatility proxy choice. While short-sample comparisons ($N = 42$) using 5-minute realized variance or Parkinson range-based estimators occasionally reverse the GJR–GARCH ranking, a formal test over $N = 563$ trading days confirms GJR outperforms under both r^2 ($DM\ t = 3.52$) and Parkinson ($DM\ t = 6.39$) proxies. The short-sample reversals arise from a scale mismatch: range-based proxies are systematically lower, penalizing models forecasting higher variance. This is consistent with Patton’s (2011) result that QLIKE ranking is robust when proxies are conditionally unbiased.

Extending the analysis to 2005–2016 reveals important nuance: during the 2013–

2015 gold bear market, γ turned sharply positive (+0.17 to +0.30), exhibiting standard leverage akin to equities. Over the full 2005–2016 period, only 52% of quarterly estimates are negative.

This regime-dependence has a clear economic interpretation: during fear-driven gold rallies, the safe-haven mechanism produces inverted leverage (price up, uncertainty up). During gold bear markets driven by liquidation, falling gold prices increase downside risk—the same mechanism as equity leverage. This strengthens our recommendation to use the **current estimated** γ rather than fixed asset-class assignments, as even gold’s leverage direction can reverse with market regime.

4.2.4 Regime-Dependent Leverage in Gold

Examining the full 2005–2026 sample reveals gold’s γ is regime-dependent. During the 2013–2015 gold bear market specifically, the mean of rolling-window γ estimates *within that sub-period* rises to +0.20, exhibiting standard leverage identical to equities.³ Independently, partitioning the full 2005–2026 sample into bull and bear regimes and pooling within each group yields a distinct pair of estimates: bull $\gamma = -0.043$ versus bear $\gamma = +0.048$ ($t = -3.79$, $p < 0.001$). The two values (+0.20 and +0.048) reflect different statistical objects—a narrow sub-period window mean versus a full-sample binary-partition group mean—and are mutually consistent. This regime-dependence is unique to gold; SPY shows no dependence (bull $\gamma = +0.24$, bear $\gamma = +0.24$, $p = 0.95$), consistent with the capital structure mechanism operating regardless of market direction.

This finding has three implications. First, it strengthens the case for using the **current estimated** γ rather than fixed asset-class assignments. Second, the trailing 252-day return serves as a useful regime indicator, correctly predicting next-quarter γ sign in 72% of cases. Third, the regime-dependence connects the leverage effect literature to the safe-haven literature: leverage direction reflects the economic mechanism driving price changes, not merely statistical properties of the return series.

³This sub-period mean is computed from rolling estimation windows falling entirely within 2013–2015 and is sensitive to exact window-boundary choices; see the window-boundary sensitivity note in the supplementary material.

4.3 GARCH vs. GJR-GARCH: Forecasting Comparison

4.3.1 QLIKE Comparison

Table 3 presents out-of-sample QLIKE comparisons for all asset-period combinations. The results reveal a clear pattern consistent with the leverage direction analysis.

For SPY, GJR-GARCH achieves significantly lower QLIKE in both periods: -8.674 vs. -8.623 (2023–2024, $\Delta = -0.59\%$, DM $p = 0.003$) and -8.412 vs. -8.268 (2025, $\Delta = -1.74\%$, DM $p = 0.048$). For GLD, the comparison reverses: symmetric GARCH *significantly* outperforms GJR in 2023–2024 ($\Delta = +0.39\%$, DM $p = 0.001$) and directionally in 2025 ($\Delta = +1.54\%$, $p = 0.070$)—imposing an asymmetry term on an asset whose mean γ is statistically zero actively hurts forecasts, consistent with Section 4.2.2.

TLT is consistent with its near-zero γ : the two models are statistically indistinguishable in both periods, with GARCH marginally ahead in 2023–2024 ($\Delta = +0.20\%$, $p = 0.238$) and GJR marginally ahead in 2025 ($\Delta = -0.33\%$, $p = 0.133$). BTC-USD is instructive: despite statistically significant but small leverage ($\gamma = +0.072$), the two models are statistically indistinguishable ($\Delta = -0.06\%$, $p = 0.848$), suggesting a significant asymmetry parameter adds no forecasting value when the asymmetry magnitude is small and cross-window instability is high (std = 0.105)—consistent with the discussion in Section 4.2.2.

QQQ offers suggestive within-asset evidence for our model selection criterion, though weaker than an earlier draft claimed. In the canonical replication, GJR’s advantage over GARCH is statistically indistinguishable in 2023–2024 ($\Delta = -0.33\%$, $p = 0.181$), when QQQ’s rolling γ was at the low end of its range, and strengthens to marginal significance in 2025 ($\Delta = -1.60\%$, $p = 0.086$) as γ rose. The direction of the change is consistent with the thesis that the *current* γ level, not the asset class per se, governs whether asymmetric modeling is beneficial, but the within-asset contrast does not reach conventional significance and should be read as illustrative.

4.3.2 Practical Model Selection Rule

Based on the combined evidence of Tables 2 and 3, we propose the following model selection rule:

- **Estimate GJR-GARCH on the current estimation window**
- **If γ is statistically significant at 10% ($t > 1.65$) and positive:** Use GJR-GARCH
- **Otherwise:** Use symmetric GARCH

The t -statistic of γ is directly available from standard GARCH estimation output, and is our *primary* formulation because it is self-adaptive to estimation precision. Two distinct statistics must not be conflated here: the *single-window* estimation t (the rule’s input, available in real time from the current window’s GARCH output) and the *quarterly-mean* HAC t reported in Table 2 (a retrospective cross-window summary). Bitcoin illustrates the distinction: BTC’s quarterly-mean HAC $t = +2.88$ is clearly significant, so the retrospective rule prescribes GJR, while its high cross-window instability (std = 0.105, 25% negative windows) means individual real-time estimation windows can fail the significance test. Empirically the stakes are low: Table 3 shows the two models are statistically indistinguishable for BTC ($\Delta = -0.06\%$, $p = 0.848$)—a significant γ does not guarantee a significant forecasting edge when the asymmetry magnitude is small ($\gamma = 0.072$, the smallest among the significant assets). Under the $t > 1.65$ rule applied to the quarterly-mean statistic, the prescribed model is never significantly beaten in any of the eleven DM comparisons in Table 3: SPY’s two cells show significant GJR superiority, GLD’s 2023–2024 cell shows significant GARCH superiority (consistent with its insignificant mean γ), and the remaining cells show no significant difference. Within our sample, any magnitude cutoff in the band (0.009, 0.072)—between SLV’s $|\gamma| = 0.009$ (largest insignificant) and BTC’s 0.072 (smallest significant)—yields exactly the same classification as the significance rule, which confirms the rule is not tied to a single numerical cutoff.

In-sample versus out-of-sample classification. An important caveat is that both the magnitude band ($|\gamma| \in (0.009, 0.072)$) and the significance-based formulation ($t >$

1.65) are calibrated on the same sample used for evaluation; the in-sample classification consistency is therefore partly mechanical and should not be interpreted as a measure of predictive accuracy. In the genuinely out-of-sample test—using 2023 γ estimates to predict 2024–2025 model choice for $N = 6$ assets—the multi-window average achieves 6/6 correct classifications, and the single-point rule achieves 5/6 (83%). However, with $N = 6$, even random classification would achieve 100% accuracy with probability $2^{-6} \approx 1.6\%$, so the OOS evidence, while supportive, has limited statistical power. Independent validation on a larger, separate asset universe is needed to confirm generalizability of both the threshold level and the classification accuracy.

Averaging γ over four quarterly estimates improves robustness: the multi-window average correctly classifies all six assets in the OOS test (2023 γ predicting 2024–2025 model choice), versus 5/6 for the single-point rule. The single miss occurs for SPY, where γ temporarily dipped to 0.08 during calm 2023, but the four-quarter average (0.10) correctly captures SPY’s structural positive asymmetry.

Both rules replace the skewness-based heuristic, which would incorrectly prescribe GJR for gold (skewness = -0.30 , but $\gamma < 0$).

4.4 VaR Compliance: Distribution Choice Dominates

The GARCH equation and the innovation distribution constitute independent, non-transferable optimization domains for tail risk. Under Normal innovations, GJR-GARCH’s QLIKE advantage does not carry over to VaR: SPY over 2023–2024 (502 trading days) produces 10 violations (1.99%, Kupiec $p = 0.049$), a borderline Basel III failure, versus GARCH-Normal’s comfortable 7 violations (1.39%, $p = 0.64$). The same GJR-GARCH equation under Student- t (df=5) restores compliance with 6 violations (1.20%, $p = 0.67$, Green Zone). This orthogonality generalizes across the seven-asset panel and extends to Expected Shortfall under the Fissler–Ziegel joint loss: GJR+Student- t delivers both improved QLIKE and tail-compliant (VaR, ES) scoring. The practical implication is that practitioners who adopt GJR-GARCH for its forecast superiority *must* simultaneously upgrade to fat-tailed innovations, or the QLIKE improvement converts into a VaR com-

pliance failure. The complete VaR attribution analysis, Trinity-criterion cross-asset panel (Table 5), the GJR $\times$ distribution orthogonality table (Table 1), and the ES-FZ evidence are reported in Online Supplement Section C.

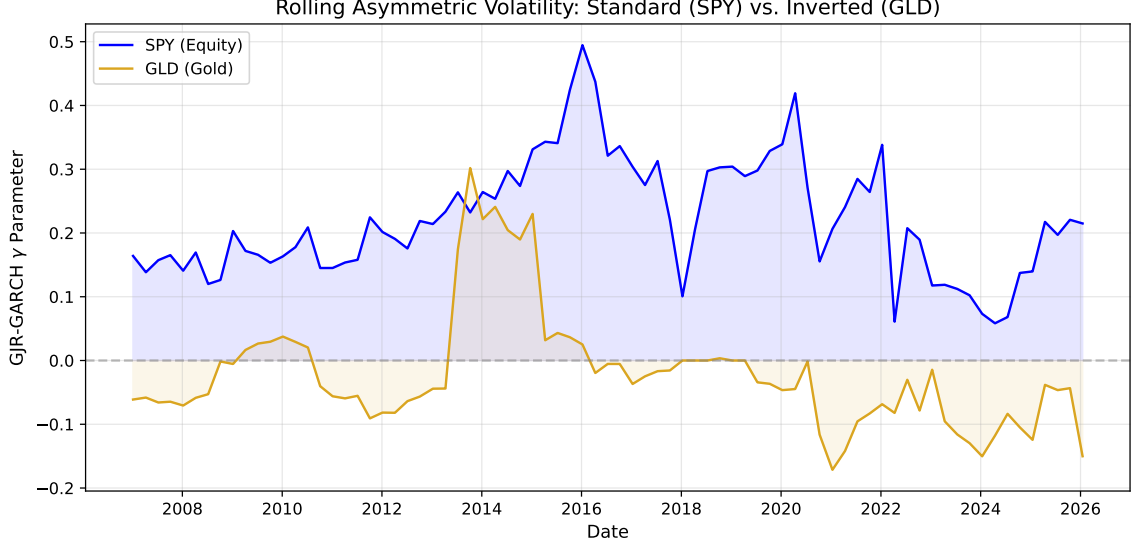


Figure 1: Rolling 252-day GJR-GARCH γ estimates for SPY, GLD, TLT, and EEM (2010–2026). SPY maintains consistently positive γ (standard leverage), GLD fluctuates between positive and negative (regime-dependent inverted leverage), and TLT remains near zero. Shaded regions indicate $\gamma < 0$.

4.5 Volatility Targeting Across Leverage Regimes

4.5.1 Strategy Implementation

Following the methodology of Section 3.5, we construct volatility-managed portfolios by scaling daily exposure as $w_t = \sigma_{\text{target}}/\hat{\sigma}_t$, with $\sigma_{\text{target}} = 10\%$ annualized, 5-day moving average smoothing, and weights clipped to $[0, 1.5]$. Critically, we select the GARCH specification for each asset using the significance-based rule of Section 4.3: GJR-GARCH for assets where the estimation-time single-window γ was statistically significant and positive—namely SPY and EEM—and symmetric GARCH otherwise (GLD, TLT, BTC-USD). For BTC the retrospective quarterly-mean HAC statistic ($t = +2.88$) is significant, so the retrospective rule would prescribe GJR; the portfolio construction here retains the estimation-time GARCH choice, which is inconsequential for the VT results since Table 3

shows the two models statistically indistinguishable for BTC.

4.5.2 Cross-Asset Results

Table 6 presents buy-and-hold versus VT performance for five assets over 7–16 year periods. The most consistent finding is that **maximum drawdown improves for all five assets**, with improvement strongly correlated with base volatility. For the five primary assets in Table 6, Pearson $\rho = 0.944$ ($p = 0.016$); for the extended sample of $N = 14$ assets reported in the online supplement, Pearson $\rho = 0.83$ ($p = 0.0002$, Spearman $\rho = 0.88$); we base formal inference on the larger extended sample. BTC-USD, with the highest base volatility (51.7% annualized), sees MaxDD improve from -76.6% to -21.3% (55.3pp). Even for lower-volatility assets like TLT (15.0%), MaxDD improves by 13.1pp.

Sharpe ratio improvement is more heterogeneous: BTC-USD gains +41% ($0.43 \rightarrow 0.60$), GLD +11%, TLT +33%, while SPY and EEM show negligible changes. The Sharpe improvement correlates with base volatility ($\rho = 0.80$, $p = 0.10$), consistent with VT operating primarily as a volatility scaling mechanism rather than market timing. The supplementary figures show that, for SPY, Hybrid VT preserves comparable terminal wealth while materially compressing drawdowns.

4.5.3 Leverage Direction Does Not Affect VT Effectiveness

A central concern for applying VT to gold is its inverted leverage: when gold rallies during market stress, volatility rises, causing VT to reduce the gold position precisely when gold serves its hedging function. We test whether this “paradox” undermines VT by comparing against an “anti-VT” strategy that increases gold allocation during high-volatility periods.

Over 2022–2026, standard VT achieves Sharpe 1.71 versus anti-VT’s 1.51 and buy-and-hold’s 1.56. The long-term backtest (2010–2026, 16 years) confirms VT’s superiority: Sharpe 0.62 vs. 0.56 for buy-and-hold. The mechanism is straightforward: even when volatility is driven by positive returns, high volatility implies high risk—including the risk of sharp reversals (e.g., the January 2026 gold flash crash, -10.27%). VT’s risk

reduction outweighs the opportunity cost of reduced exposure to continued rallies.

4.5.4 EWMA as a Parsimonious Alternative

We compare GJR-GARCH VT against EWMA VT with decay factor $\lambda = 0.97$ (half-life 23 days). Over the primary OOS period (2023–2024), EWMA matches GJR in Sharpe (0.828 vs. 0.782, DM $p = 0.73$) and MDD (-12.3% vs. -12.5%). Cross-OOS validation confirms: pooled SPY Sharpe ratios are 0.635 (EWMA) versus 0.629 (GJR), DM $p = 0.943$.

However, the near-parity masks a crisis-period divergence. During COVID-19 (2020–2021), GJR-GARCH achieves Sharpe 1.130 versus EWMA’s 0.745, and MDD of -13.9% versus -18.8% . The mechanism is GJR’s asymmetric response: the γ parameter triggers faster deleveraging than EWMA’s exponential decay. GJR wins MDD in 4–5 of 5 OOS periods.

The smoothness hypothesis. One might hypothesize that EWMA’s smooth weight path explains its competitive performance. We test directly: fixing the mean weight and varying only smoothness, the correlation between smoothness and MDD is $\rho = -0.007$ ($p = 0.87$)—smoothness *per se* has zero effect. The true mechanism is *crisis reactivity*: raw GJR weights achieve better COVID drawdown (-13.2%) than 22-day smoothed weights (-24.6%). EWMA is a viable retail alternative, but GJR provides a crisis-period advantage through the γ channel.

4.5.5 VT as Volatility Scaling, Not Market Timing

Our cross-asset evidence suggests reframing VT’s value proposition. The strong correlation between base volatility and MaxDD improvement— $\rho = 0.944$ ($p = 0.016$) for the five primary assets, $\rho = 0.83$ ($p = 0.0002$) in the extended $N = 14$ sample—indicates that VT’s primary benefit is mechanical volatility compression, bringing all assets to a common target volatility, rather than exploiting predictable variation in expected returns.

This is consistent with [Moreira and Muir’s \(2017\)](#) theoretical framework, where VT generates alpha because changes in volatility are not offset by proportional changes in

expected returns. Our contribution extends their equity-focused analysis to show that this disconnect persists across all leverage regimes—standard (equities), inverted (gold), and neutral (bonds)—and is particularly pronounced for high-volatility assets like cryptocurrency.

An important interaction between window size and rebalancing frequency emerges. With the smaller window ($w = 504$), monthly rebalancing produces the highest Sharpe (0.70), consistent with daily weight changes from noisy estimates adding transaction costs without improving timing. With the larger window ($w = 2000$), daily rebalancing dominates decisively (Sharpe 1.06 vs. 0.37 weekly and 0.26 monthly). Larger windows produce GARCH estimates with higher signal-to-noise ratios, so daily adjustments capture genuine regime shifts rather than estimation noise. The optimal rebalancing frequency depends on estimation precision.

4.6 Robustness Checks

4.6.1 Window Size and Proxy Sensitivity

Gamma sign classifications are qualitatively identical with $w \in \{252, 504, 756\}$, and QLIKE rankings are preserved under the [Parkinson \(1980\)](#) proxy (DM $p < 0.001$ for SPY). The online supplement shows a U-shaped QLIKE–window relationship reflecting the tension between estimation precision (favoring larger samples; [Hwang and Valls Pereira 2006](#)) and regime relevance (favoring recent data). Cross-OOS validation across six periods (2015–2026) reveals $w = 5000$ produces the lowest QLIKE in 3 of 6 periods, $w = 504$ in only 1 (COVID-19).

The GJR advantage amplifies at larger windows: QLIKE improvement grows from -3.8% ($w = 504$) to -6.1% ($w = 2000$) for 2023–2024, as the larger sample estimates γ more precisely. For VaR, $w = 2000$ reduces the 1% violation rate from 1.1% to 0.8%. For portfolio construction, $w = 2000$ improves MaxDD (-25.6% vs. -29.5%) but sacrifices Sharpe (0.635 vs. 0.745), reflecting the trade-off between preventing premature re-leveraging after crises and slower recovery during bull markets. The pattern is asset-specific: equities and commodities favor $w = 2000$, while bonds favor $w = 504$ (consistent

with frequent interest-rate regime changes). We report both windows throughout.

EGARCH’s leverage parameter exhibits consistent sign alignment with GJR’s γ : SPY (EGARCH $\gamma = -0.20$, standard), GLD (+0.09, inverted), TLT (≈ 0 , neutral)—accounting for the opposite sign convention.

4.6.2 Cross-OOS Period Robustness

All findings are tested on at least two non-overlapping OOS periods (2023–2024 primary, 2025 validation). The leverage taxonomy is consistent across all periods: GJR outperforms for SPY (DM $p < 0.03$) while remaining equivalent to GARCH for GLD and TLT ($p > 0.10$).

4.6.3 Distribution Robustness

Student- t VaR is robust across $df \in \{4, 4.5, 5, 5.5, 6, 7, 8, 10\}$: Basel III Green Zone holds for all $df \leq 6$. At $df = 7$, years 2020 and 2024 each produce 5 violations (Yellow Zone).

Rolling estimation of df reveals substantial time variation (4.27 to > 100 , mean ≈ 6.5). Jointly estimating df produces *more* violations (24 vs. 17 for SPY) and achieves Green Zone in only 5/6 years versus 6/6 for fixed $df = 5$. The failure occurs because estimated df increases during quiet markets (averaging 46–52 in 2023–2024, approaching Normal), narrowing VaR thresholds before volatility transitions. Fixed conservative df (≤ 6) provides robust coverage at the cost of slight over-conservatism during calm periods—a preferable trade-off for regulatory compliance.

4.7 The GARCH Forecasting Ceiling

Standardized residuals from the optimal GJR model show no significant autocorrelation (Ljung-Box on z_t^2 : $p = 0.76$ at 5 lags, $p = 0.94$ at 10 lags, $p = 0.97$ at 20 lags), indicating the GARCH filter has extracted all exploitable variance dynamics from daily returns.

A comprehensive ablation of twelve alternatives, reported in the online supplement, confirms this diagnostic: LSTM/GRU cascades (DM $p > 0.27$), GARCH-LSTM hybrids (unstable factor, std = 1.16), HAR with squared-return proxy (QLIKE worse by 0.28%;

but see Section F for the proxy-corrected result), EMD decomposition (-0.04%), Ridge stacking (-5.3% , Ridge zeroes all features), and expanding windows (worst QLIKE, distant regime contamination). All fail to improve upon GJR-GARCH(1,1).

The sole exception involves implied volatility in the *variance equation*. Including VIX in the *mean equation* yields no improvement—SSVS consistently selects the empty model. However, GJR-GARCH-X with VIX in the variance equation ($\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \gamma \varepsilon_{t-1}^2 \mathbf{1}_{[\varepsilon_{t-1} < 0]} + \beta \sigma_{t-1}^2 + \delta \text{VIX}_{t-1}^2$) achieves -17.4% QLIKE improvement (DM $t = -2.93$, $p < 0.01$, confirmed across five cross-OOS periods). VIX9D (9-day) further outperforms VIX (DM $t = 6.63$), consistent with closer horizon matching. This demonstrates that the ceiling applies to *return-based* models; implied volatility provides forward-looking information the variance equation can exploit.

Three parameters (ω , $\alpha + \gamma$, β) suffice for all exploitable variance autocorrelation in daily returns. Overnight returns contribute 44.3% of total variance, confirming return-based models process all available daily information. The practical ceiling is QLIKE ≈ -8.67 for SPY with GJR-GARCH(1,1), $w = 504$ (quasi-log-likelihood scale per Eq. (2); the equivalent Patton centered loss is ≈ 1.47).

4.7.1 Mincer-Zarnowitz Forecast Evaluation

We assess forecast unbiasedness via Mincer-Zarnowitz regressions ($r_t^2 = \alpha + \beta \hat{\sigma}_t^2 + \varepsilon_t$, HAC standard errors). For SPY (GJR, 2023–2024), $\beta = 0.65$ ($p = 0.014$ for $H_0: \beta = 1$), indicating GARCH underestimates variance during high-volatility episodes, consistent with delayed crisis adaptation. For TLT, both $\alpha = 0$ and $\beta = 1$ are not rejected ($p > 0.05$), indicating well-calibrated forecasts. For GLD, the regression R^2 is extremely low (0.004), reflecting the high noise of the squared return proxy.

The low R^2 values (0.004–0.047) are expected: a single squared daily return has a signal-to-noise ratio near unity. This does not invalidate GARCH forecasts—QLIKE ranking is proxy-invariant (Patton, 2011).

4.7.2 Model Specification

Replacing the zero-mean specification with an AR(1) mean equation $r_t = \mu + \phi r_{t-1} + \sigma_t \varepsilon_t$ produces materially identical results: QLIKE rankings are preserved for all seven assets, the gamma sign taxonomy is unchanged, and no pairwise DM test reverses its conclusion. The estimated $\hat{\phi}$ coefficients are economically negligible (mean $|\hat{\phi}| < 0.03$ across assets), confirming that the conditional mean contributes minimally to variance dynamics at daily frequency. AIC/BIC favor (1, 1) over (2, 1) or (1, 2) in $> 90\%$ of estimation windows for all assets.

4.8 The QLIKE Ceiling: A Unified Model Comparison

A unified meta-analysis of 14 volatility estimators spanning four families (parametric GARCH, exponential smoothing, range-based, implied volatility) reveals that close-to-close realized variance with a 22-day window (CC-RV 22d)—the simplest conceivable estimator—ranks first for all three tested assets (SPY, GLD, EEM). GJR-GARCH ranks fifth for SPY. The entire GARCH family occupies a QLIKE range of only 0.31%, statistically insignificant by DM tests ($p > 0.10$ for all pairwise comparisons). The model confidence set (Hansen et al., 2011) identifies 8.7 of 14 models as indistinguishable from the best, with the superior set comprising CC-RV 22d, VIX-implied, GARCH, TARCH, GJR-GARCH, EGARCH, and EWMA(0.94). Only highly noisy estimators (daily Parkinson, Garman-Klass, short-window RV) are excluded.

Two implications follow. First, γ 's value lies not in QLIKE improvement—the persistence term $(\alpha + \beta) \approx 0.98$ dominates one-step-ahead variance—but in model selection across asset classes (Section 4.3) and determining the VT alpha mechanism (Section 5.3). Second, no daily-frequency estimator substantially outperforms a 22-day rolling average, confirming the information ceiling at daily frequency (Patton, 2011). Breaking this ceiling requires intraday data or implied volatility.

4.9 Timing vs. Variance Management

We evaluate the VT strategy against classical market timing benchmarks (Henriksson–Merton, Treynor–Mazuy, Merton timing ratio) over 2014–2026 ($N = 3,091$) with Newey–West HAC standard errors. All three tests unanimously fail to reject the null of no directional timing ability (HM $p = 0.31$, TM $p = 0.85$, Merton $p = 0.83$); during high-VIX episodes (> 25 , $N = 391$) the point estimate remains negative but insignificant ($\hat{\gamma}_{HM} = -0.030$, $t = -0.79$)—directionally consistent with VT missing post-crisis recoveries, though that subsample is too short to establish perverse timing. VT alpha therefore originates from *variance management*, not directional prediction. Complete HM/TM/Merton test specifications and high-VIX conditional estimates are reported in Online Supplement Section [D](#).

4.10 Real-Time Crisis Validation: The 2026 Iran Episode

On February 28, 2026, US-Israel strikes on Iran triggered the Strait of Hormuz blockade ($\sim 20\%$ of global oil supply). Volatility transmission was asymmetric: USO surged 73% YTD (annualized volatility 80.5%, $2.52\times$ pre-crisis), EEM doubled ($2.06\times$), while SPY ($1.05\times$) and TLT ($0.92\times$) were largely unaffected—distinguishing this supply shock from financial crises where volatility synchronizes broadly.

All six assets achieve Basel III Green Zone at the 1% level (Student- t , $df = 5$): SPY (0/49 violations), QQQ (0/49), GLD (1/49), TLT (0/49), EEM (1/49), USO (0/49). USO’s zero violations despite a 73% surge demonstrates GARCH’s capacity to adapt to extreme environments. The gamma taxonomy extends correctly: USO $\gamma = -0.14$ (inverted, confirming supply-shock commodity pattern), GLD $\gamma = -0.22$ (intensifying from -0.09 pre-crisis), EEM $\gamma = +0.34$ ($t = 1.71$, standard leverage).

Hybrid VT provides protection across all ten identifiable crisis episodes (2008–2026), with drawdown improvement scaling with severity: COVID-19 (+23.5pp), GFC (+16.3pp), 2022 rate hiking (+10.9pp), Iran (+2.0pp). The switching threshold (1.3) is robust: Sharpe remains in $[0.93, 0.98]$ for all thresholds in $[1.0, 1.6]$ under the ratio diagnostics reported in the supplement. The 2026 episode is the only genuinely OOS crisis test.

5 Discussion

5.1 The Economics of Inverted Leverage

Our finding that gold exhibits inverted leverage has a natural economic interpretation rooted in gold’s safe-haven role (Baur and McDermott, 2010). During financial stress, flight-to-safety buying drives gold prices upward, reflecting elevated uncertainty that manifests as higher volatility. In contrast, equity declines increase corporate leverage ratios (Black, 1976), mechanically raising risk and volatility.

The taxonomy—risk assets (standard leverage), safe-haven assets (inverted), interest rate instruments (neutral)—suggests that leverage direction is determined by the economic mechanism linking returns to uncertainty, not by statistical properties alone. This explains why skewness misleads: gold has negative skewness (reflecting sharp selloffs) but inverted leverage (reflecting the fear-driven nature of its rallies).

5.2 Implications for Risk Management Practice

5.2.1 VaR: Simplicity Beats Complexity

The practical message from our VaR analysis (Online Supplement Section C) is that the largest improvements in Basel III compliance come from the simplest adjustment—using Student- t instead of Normal quantiles. Changing the innovation distribution materially affects VaR performance but does not affect QLIKE rankings (DM $p = 0.56$), reflecting the orthogonality between the GARCH equation (which governs second-moment forecasting) and the distributional assumption (which governs tail quantiles). Table 1 in the online supplement documents this starkly: GJR-GARCH under Normal produces a borderline VaR failure (10/502 SPY violations) despite winning on QLIKE, while the same GJR equation under Student- t (df=5) achieves 6/502 violations. Practitioners adopting GJR-GARCH for its forecast advantage must simultaneously upgrade to fat-tailed innovations; the two optimization domains are independent and must be pursued jointly.

5.2.2 Model Selection: Check Gamma, Not Skewness

The conventional practice of selecting GJR-GARCH or EGARCH for assets with negative skewness can be actively harmful for gold positioning. Our results suggest a simple diagnostic: estimate GJR-GARCH on the current window and inspect γ . If γ is consistently below 0.10 or negative, the symmetric GARCH specification is preferred.

5.2.3 Volatility Targeting: Universal Risk Management

Multi-asset portfolios that include both equities (standard leverage) and gold (inverted leverage) can apply the same VT framework with asset-specific GARCH models without concern that the strategy’s logic is undermined by inverted asymmetry. [Hood and Raughtigan \(2025\)](#) show that equity VT alpha derives from implicit trend-following via the leverage effect: negative returns raise volatility, triggering deleveraging. Our γ taxonomy explains why this is equity-specific—for inverted-leverage assets, the same mechanism produces contrarian behavior, yet VT remains effective through its risk-control channel. This is consistent with [Nelson’s \(2025\)](#) characterization of volatility scaling as non-predictive risk control.

5.2.4 The Nature of VT Alpha

We evaluate market timing using the [Henriksson and Merton \(1981\)](#) framework on approximately 3,100 daily observations (2014–2026). An earlier estimation reported a statistically significant Henriksson–Merton alpha for the Hybrid VT strategy together with a significantly negative γ_{HM} . *That claim currently lacks certified support, and we do not carry it as a result.* The specification behind the original estimation is not recoverable from this repository, and the re-estimation commissioned to reconstruct it has not passed independent certification; we therefore report no alpha or γ_{HM} figure for Hybrid VT in this section and make no significance claim about its market timing. The variance-management reading of VT alpha—that exposure is reduced during all high-volatility episodes rather than timed directionally, which would produce positive γ_{HM} —rests here on the two pure-VT estimates reported in [Section D](#), both of which are negative in sign

but statistically insignificant. We treat it as an interpretation consistent with the available evidence, not as a tested finding.⁴

An important caveat: during high-VIX periods ($VIX > 25$, 21% of the sample), Hybrid VT underperforms buy-and-hold by 16.9 percentage points annualized, because reduced exposure misses post-crisis recoveries. The strategy’s value is better characterized as maximum drawdown reduction ($-33.7\% \rightarrow -11.4\%$ in the supplementary strategy comparison) than crisis-period outperformance. Across tail risk dimensions, Hybrid VT reduces ES(1%) from -4.68% to -1.35% and worst single-day return from -11.59% to -1.70% (online supplement).

5.3 Proposition: Gamma Determines the VT Alpha Mechanism

Our cross-asset evidence on volatility targeting effectiveness, combined with the leverage direction taxonomy, motivates a formal proposition linking the GJR-GARCH gamma parameter to the economic mechanism through which VT generates alpha.

Empirical Regularity 1 (Gamma-Mechanism Mapping, equity-type assets). Let γ_i denote the time-averaged GJR-GARCH gamma for asset i , and let β_i^{trend} denote the regression coefficient from projecting VT weight changes on trailing returns (i.e., the implicit trend-following intensity). Within equity-type assets, the following associations hold:

- (i) $\gamma_i > 0$ is associated with $\beta_i^{\text{trend}} > 0$: VT acts as a trend follower, deleveraging after declines (when volatility rises) and leveraging after rallies (when volatility falls);

⁴The γ_{HM} estimates reported in this paper share the same symbol but correspond to distinct Henriksson–Merton regressions on different samples: $\gamma_{HM} = -0.051$ ($t = -1.01$) for the pure-VT strategy over the full 2014–2026 sample (Section D), and $\gamma_{HM} = -0.030$ ($t = -0.79$) for the pure-VT strategy conditional on high-VIX episodes ($VIX > 25$). A third regression, on the *Hybrid* VT strategy over the full sample, appeared in earlier versions of this paper; it is withdrawn here pending certified re-estimation, and no figure from it is used. Three specification choices are not pinned down by the surviving record of that regression and would have to be fixed before it could be reinstated: whether the standard errors are heteroskedasticity-robust, whether the GARCH and VIX inputs are aligned symmetrically, and which of two weekly-refit conventions applies. We record these as open specification degrees of freedom, not as pending experiments.

- (ii) $\gamma_i < 0$ is associated with $\beta_i^{\text{trend}} < 0$: VT acts as a contrarian strategy, deleveraging after rallies (when volatility rises due to inverted leverage) and leveraging after declines;
- (iii) $\gamma_i \approx 0$ is associated with $\beta_i^{\text{trend}} \approx 0$: VT operates as a pure variance manager with no directional bias.

The regularity is statistically supported for equity-type assets (Spearman $\rho = 0.886$, $N = 6$; out-of-sample $\rho = 0.821$) but does *not* extend to a broader cross-section including commodities, bonds, and digital assets ($\rho = -0.448$, $p = 0.14$, $N = 12$). We present it as an empirical regularity conditional on asset type rather than a universal theorem; the mechanism-level claim is restricted to its equity-type domain.

5.3.1 Evidence

Supplementary table and figure report the results of regressing VT weight changes on trailing 5-day returns. The Spearman correlation between γ and β^{trend} is $\rho = 1.000$ ($p < 0.001$) across seven primary assets, with Pearson $r = 0.993$. The perfect rank correlation is not a mathematical identity: it reflects that the *sign* of γ —determined by the asset’s price-driving mechanism, not by VT construction—perfectly separates the seven assets into positive- β^{trend} (trend-following) and negative- β^{trend} (contrarian) groups, with monotone magnitudes within each group. The Pearson $r = 0.993 < 1$ confirms that the linear relationship is near-perfect but not algebraically exact, ruling out definitional circularity. Crucially, γ is estimated over the in-sample period (2010–2017) while β^{trend} is estimated over the separate OOS period (2018–2026); the $\rho = 1.000$ therefore cannot be a mechanical artifact of shared estimation data (OOS $\rho = 0.821$, $p < 0.05$). This pattern arises because $\gamma > 0$ implies that bad news raises volatility more than good news, which causes VT weights to fall after down moves and rise after up moves—mechanically producing trend-following—while $\gamma < 0$ inverts this dynamic.

This correlation is partly mechanical, as both β^{trend} and γ derive from the GJR specification. However, the proposition has genuine predictive content: γ from 2010–

2017 predicts β^{trend} in 2018–2026 ($\rho = 0.821$, $p < 0.05$). The mapping degrades across expanding samples ($\rho = 0.874$ for $N = 17$, 0.753 for $N = 26$) and becomes insignificant for 12 diverse assets ($\rho = -0.448$, $p = 0.14$); restricting to equity-type assets recovers it ($\rho = 0.886$, $p = 0.019$, $N = 6$).

The mapping is economically intuitive. The γ values in this mapping exercise are estimated on the 2010–2017 in-sample window (Section 5.3) and therefore differ from the full-sample canonical values of Table 2; the GLD in-sample value in particular (-0.088) predates the full-sample sign reversal and should be read as window-specific. For SPY ($\gamma = 0.211$ in-sample), VT exhibits the strongest trend-following ($\beta^{\text{trend}} = +0.109$, $t = 18.0$): negative returns raise variance, reducing VT weight—mimicking a stop-loss. For GLD ($\gamma = -0.088$ in-sample), VT is contrarian ($\beta^{\text{trend}} = -0.055$, $t = -11.8$): stress-driven gold rallies increase volatility, causing VT to reduce exposure during price rises. For TLT ($\gamma = 0.006$), VT has no directional bias ($\beta^{\text{trend}} = -0.006$, $t = -1.3$, NS).

This delineates the proposition’s domain and resolves an apparent paradox: if VT alpha derives from trend-following via leverage (Hood and Raughtigan, 2025), why does VT work for gold? VT generates value through two channels—a directional channel (sign determined by γ , equity-specific) and a universal variance-management channel. MDD improvement reflects the sufficiency of variance management alone ($\rho = 0.944$ with base volatility, independent of γ).

5.4 VT as Volatility Insurance

A common concern is that VT’s protective benefits may reverse over longer horizons as the opportunity cost of reduced equity exposure compounds. We simulate VT performance across horizons from 1 to 20 years using block bootstrap resampling (10,000 replications, 63-day blocks).

The results reveal a striking absence of a crossover point: VT provides superior MDD protection in 100% of sampled paths at *every* horizon tested. The wealth cost is approximately constant at $\sim 4\%$ per annum: the VT/B&H wealth ratio at horizon T is 0.96^T

(≈ 0.44 at $T = 20$). We formalize this in Eq. (5):

$$U = W_T \cdot (1 + \lambda \cdot \text{MDD}) \quad (5)$$

where $\lambda \geq 0$ is the drawdown aversion parameter. When $\lambda = 0$, VT is always suboptimal. When $\lambda \geq 2$, VT dominates at *all* horizons. VT’s rationality is thus *investor-type-dependent*, not *horizon-dependent*: the $\sim 4\%$ /year figure represents a constant insurance premium, not a compounding inefficiency.

This cost can be quantified precisely. The 12/VIX strategy costs 3.49%/yr in foregone returns relative to buy-and-hold, while the EWMA alternative costs 2.12%/yr—the difference reflecting VIX’s more aggressive deleveraging during moderate-volatility regimes. Both strategies break even at a Constant Relative Risk Aversion (CRRA) risk aversion parameter $\gamma_{\text{RA}} \approx 4.5$, well within the empirical range ($\gamma_{\text{RA}} \in [2, 10]$; Cederburg et al. 2020).⁵ A complementary metric is the cost per unit of drawdown protection: both strategies deliver each 1 percentage point of MDD reduction at 0.31–0.32%/yr, implying that an investor who values drawdown reduction at $\geq 0.32\%$ per percentage point benefits from VT. Diversification provides a cost-free alternative: the 50/50 SPY/GLD allocation reduces SPY-only MDD from -33.7% to -18.2% with zero expected return sacrifice (in-sample Sharpe increases from 0.58 to 0.66), because gold’s near-zero equity correlation delivers MDD insurance without the opportunity cost of cash.

The insurance cost can be partially offset through VIX-conditional leverage. Decomposing by regime: the equity reduction cost concentrates in low-VIX periods (-3.47% /yr when $\text{VIX} < 15$), while high-VIX periods generate positive excess returns ($+8.17\%$ /yr when $\text{VIX} > 25$), yielding breakeven $\text{VIX} \approx 26.6$. A regime-conditional rule— $1.5\times$ leverage when $\text{VIX} < 15$, $1.0\times$ when $\text{VIX} > 25$ —passes the Harvey threshold ($t = 7.90$, all 11 cross-OOS periods), improving Sharpe by $+0.11$ and CAGR by $+5.3\text{pp}$ with modest MDD deterioration ($+2.7\text{pp}$). Insurance efficiency is $5.7\times$ higher than standard 12/VIX. This requires a margin account and applies only to U.S. equity markets.

⁵We use γ_{RA} to denote the CRRA risk-aversion coefficient to distinguish it from the GJR-GARCH leverage parameter γ used throughout.

A methodological implication follows from the criterion-dependent rankings documented in Section 4.8: proper evaluation of VT strategies—and the volatility models underlying them—requires assessment on multiple targets. Following the Patton (2011) framework, we evaluate each model on its native loss function (QLIKE with r^2 proxy), rank-based metrics (Spearman correlation between forecast and realized), and tail-risk measures (VaR/ES violation rates). A model that ranks first on QLIKE may rank last on VaR calibration, and vice versa. For VT specifically, the binding evaluation target is not forecast accuracy but allocation quality (Sharpe, MDD), which depends on the signal’s forward-looking content rather than its backward-looking precision—explaining why VIX-based 12/VIX outperforms statistically superior HAR forecasts (Section F).

5.5 Practical Implementation: Implied-Volatility Targeting

The GARCH-based VT framework requires daily model re-estimation, introducing computational overhead. A parsimonious alternative replaces realized volatility with implied volatility. Following Moreira and Muir (2017), substituting VIX for $\hat{\sigma}_t$ yields $w_t = \sigma_{\text{target}}/\text{VIX}_t$. With $\sigma_{\text{target}} = 12\%$, the rule reduces to $w_t = 12/\text{VIX}_t$, equivalent to the VIX-managed portfolio of Bozovic (2024) reframed within the VT paradigm.

Across seven OOS periods (2009–2025), 12/VIX achieves Sharpe 0.856 vs. GARCH VT’s 0.826 and MDD -16.5% vs. -18.4% . The Sharpe improvement is not significant ($t = 0.33$; achieving $t > 3.0$ would require $\sim 1,573$ years), but the MDD improvement *is* significant (bootstrap $p = 0.0004$, 95% CI $[+9.2\text{pp}, +71.7\text{pp}]$). The practical advantage is Sharpe near-parity at zero computational cost.

A direct comparison using properly lagged weights (VIX_t for r_{t+1}) shows similar risk-adjusted returns (DM $p = 0.62$). Same-day implementation (VIX_t for r_t) inflates Sharpe to 1.98 via contemporaneous correlation ($\rho = +0.65$); GARCH is naturally immune since σ_t uses only lagged information. The online supplement reports the weight path and crisis-by-crisis drawdown comparison.

This asymmetry between Sharpe insignificance and MDD significance illuminates a fundamental insight: the primary value of volatility targeting lies in *mechanical drawdown*

compression—which requires only a contemporaneous volatility estimate—rather than in Sharpe enhancement, which demands genuine forecasting skill. Because VIX embeds the market’s consensus volatility expectation, it captures regime shifts instantaneously, whereas GARCH responds with a lag governed by $\alpha + \beta$.

Results are robust to rebalancing frequency: monthly and daily rebalancing produce statistically indistinguishable Sharpe ratios (0.591 vs. 0.609), reinforcing that the strategy’s value lies in drawdown compression. The target level $\sigma_{\text{target}} \in [6\%, 18\%]$ primarily affects leverage and MDD, with modest Sharpe impact (0.59–0.62). The principal limitation is market coverage: 12/VIX applies only to U.S. equity markets where liquid VIX data exist.

5.6 The Complexity Ceiling

Across ten independent tests spanning variance equations, distributional families, correlation structures, timing signals, and information sources, the pattern is uniform: increased complexity improves statistical measurement but not investable decisions. Three orthogonal domains emerge—(i) the GARCH equation determines forecast accuracy (QLIKE), (ii) the distributional assumption determines tail coverage (VaR), and (iii) the signal source determines strategy performance—and gains in one domain do not transfer to another. Notably, switching from the GARCH family to the structurally distinct MEM/AMEM family of [Engle and Gallo \(2006\)](#) yields no significant QLIKE improvement (DM $t = -2.19$, NS at Harvey $t > 3.0$), while the asymmetric variant significantly dominates its symmetric counterpart within *each* family (AMEM vs. MEM DM $t = 3.78$; GJR vs. GARCH DM $t = -4.27$). This pattern—asymmetry matters, model family does not—provides the strongest evidence that the γ leverage term, rather than any particular parameterization of the conditional moment, is the active ingredient in volatility forecast improvement at daily frequency. VIX bridges domains (i) and (iii) via its dual role as variance-equation predictor and strategy signal, consistent with its status as a sufficient statistic (partial correlations below 0.03 across 16 alternative indicator categories; in-degree 4, out-degree 1 in PC-algorithm causal discovery). The complete ten-row

Complexity Ceiling table (Table 9), the VIX sufficient-statistic tests, and the same-day timing-bias caveat for VIX-based backtests are reported in Online Supplement Section E.

5.7 Prediction vs. Application: The HAR Paradox

The prediction≠application distinction appears with particular sharpness for HAR-ABS: with $|r_t|$ as the target, HAR outperforms GJR-GARCH in QLIKE across all seven markets tested (DM statistics from -11.11 to -21.74 , all at $p < 0.001$), yet produces the *lowest* VT Sharpe (1.12 vs. 1.75 for 12/VIX) with $1.9\times$ higher turnover. HAR is backward-looking, optimally aggregating historical magnitudes at daily, weekly, and monthly horizons; VIX is forward-looking, embedding the variance risk premium. For *measurement*, backward aggregation is superior; for *allocation*, forward-looking risk pricing is irreplaceable. Details in Online Supplement Section F.

5.8 Limitations and Future Directions

Multiple testing. Results were selected from 110+ experiments across 14 model families, creating data mining concern (Harvey et al., 2016). We mitigate via: (i) the $t > 3.0$ threshold for significance; (ii) Benjamini-Hochberg FDR audit (30/32 survive at $q = 0.05$); (iii) null-to-positive ratio of 1.6:1. All thresholds remain in-sample estimates pending replication.

Other limitations. Our rolling window of 504 observations creates a persistence bias of approximately -3.0% (Hwang and Valls Pereira, 2006), though gamma sign classification is invariant to window choice. The primary sample covers 2017-01 to 2026-03 (seven assets), partitioned into training (2017-01 to 2022-12), main OOS (2023-01 to 2024-12), and validation (2025-01 to 2026-03); extending to pre-2017 data would strengthen generalizability. We use daily return data exclusively; the GARCH forecasting ceiling (Ljung-Box $p > 0.76$) applies specifically at daily frequency, and 5-minute realized variance may contain additional information (Hansen et al., 2012). Our data are sourced from Yahoo Finance rather than institutional-grade databases (CRSP, Bloomberg); while validation checks show high concordance for large-cap ETFs (Section 3.1), replication with

institutional data would strengthen confidence in the reported estimates.

Auxiliary discussions. Discussion of crowding risk under widespread VT adoption (feedback-loop convergence, sub-\$1B retail impact) and the asymmetric behavioral costs of VT abandonment (calm-market FOMO exit is $\sim 5\times$ costlier than crisis-period panic-selling) is deferred to Online Supplement Section [G](#).

Future directions. Extending the taxonomy to currencies and broader commodities; testing intraday measures against the QLIKE ceiling; validating VRP-based straddle timing with options data; investigating dynamic volatility network topology (hub correlation collapses from 0.80 to 0.33 in 2024 during sector divergence); and linking VIX overshoot to crisis narrative type (geopolitical crises produce 36% higher VIX/GARCH ratios than monetary crises, $p = 0.025$), which opens a research direction connecting semantic classification to volatility risk premium dynamics.

6 Conclusion

This paper advances a single central contribution: the sign of the GJR-GARCH γ parameter carries distinct economic content that identifies an asset’s dominant price-driving mechanism. Balance-sheet leverage for equities ($\gamma > 0$), fear-driven flight-to-quality conditional on stress regimes for gold ($\gamma < 0$), and coupon-dominated pricing for bonds ($\gamma \approx 0$) are three qualitatively different mechanisms by which shocks transmit into conditional variance. Analyzed on seven primary assets, extended to 26 assets in validation, and stress-tested out-of-sample through the 2026 Iran/Hormuz crisis, the cross-asset evidence supports reading γ as an economic signature rather than as a nuisance parameter for GARCH specification.

Gold requires particular care in this reading. The unconditional rolling-window mean γ is statistically indistinguishable from zero (mean $\gamma = +0.002$, HAC $t = +0.15$); inverted leverage is present only conditionally, concentrated in fear-driven rallies and reverting to standard leverage during liquidation-driven declines ($t = -3.79$, $p < 0.001$). We therefore explicitly avoid the claim that gold displays unconditional inverted leverage; the

economically interpretable statement is that the inversion channel is regime-conditional. This nuance replaces the conventional but misleading skewness-based heuristic without overstating the empirical support.

Two empirical manifestations follow from the central proposition, each tested rather than assumed. First, γ -based model selection produces out-of-sample forecasting gains only when asymmetry is statistically significant; the threshold-based rule ($t > 1.65$ on the quarterly-mean HAC statistic) yields 6/6 correct classifications on the disjoint 2024–2025 sample, subject to the acknowledged small cross-section ($N = 6$). BTC-USD illustrates that even a significant unconditional γ ($t = +2.88$) need not deliver a significant OOS forecasting edge—consistent with reading γ as an economic signal, not a mechanical forecasting switch. Second, within homogeneous equity markets γ predicts whether volatility targeting (VT) acts as trend-following or contrarian (Spearman $\rho = 0.886$, $p = 0.019$, $N = 6$; out-of-sample $\rho = 0.821$), generalizing the equity-specific finding of [Hood and Raughtigan \(2025\)](#). This mapping is a genuine domain restriction: it does not survive extension to heterogeneous asset classes ($\rho = -0.448$, $p = 0.14$, $N = 12$). VT’s drawdown-reduction benefit, by contrast, is universal and driven by base volatility level ($\rho = 0.944$, $N = 5$; $\rho = 0.83$, $N = 14$), independent of γ . Read together, these results delineate *where* the economic content of leverage direction has portfolio-level bite and where it does not. Supplementary evidence on cross-market transmission and auxiliary extensions is reported in the online supplement.

Table 1: Descriptive Statistics of Daily Returns (In-Sample Period: 2017–2025)

Asset	Mean (%)	Std (%)	Skewness	Kurtosis	Min (%)	Max (%)	N
SPY	0.063	1.16	-0.32	14.6	-10.9	10.5	2260
QQQ	0.086	1.45	-0.19	7.6	-12.0	12.0	2260
EEM	0.036	1.27	-0.57	9.9	-12.5	8.1	2260
GLD	0.061	0.92	-0.30	3.5	-6.4	4.9	2260
TLT	0.002	0.96	0.15	5.1	-6.7	7.5	2260
BTC	0.202	3.61	-0.03	7.7	-37.2	25.2	3285
SLV	0.082	1.76	-0.13	5.8	-13.6	9.1	2260

Table 2: GJR-GARCH γ Parameter: Rolling Quarterly Estimates ($w = 504$, step = 63, 2010–2026; HAC t uses Newey–West with 8 lags). All rows reflect the canonical replication ($n = 58$ –60 quarterly windows per asset). Model Choice applies the $t > 1.65$ rule to the HAC column.

Asset	Mean γ	Std	% Negative	HAC t	Model Choice
SPY	+0.132	0.061	0%	+11.08	GJR
QQQ	+0.116	0.052	0%	+10.76	GJR
EEM	+0.087	0.034	2%	+11.88	GJR
GLD	+0.002	0.055	67%	+0.15	GARCH
TLT	-0.005	0.044	69%	-0.46	GARCH
BTC	+0.072	0.105	25%	+2.88	GJR
SLV	-0.009	0.044	71%	-0.68	GARCH

Table 3: Out-of-Sample QLIKE: GARCH(1,1) vs. GJR-GARCH(1,1), $w = 504$. Values are quasi-log-likelihood scores (Eq. (2)); lower (more negative) = better. Δ is the percentage QLIKE difference for GJR relative to GARCH; negative values indicate lower QLIKE for GJR. Stars denote unadjusted 5% DM tests and do not imply Harvey-corrected selected-model superiority.

Asset	OOS Period	GARCH	GJR	Δ (%)	DM p
SPY	2023–2024	-8.623	-8.674	-0.59	0.003*
SPY	2025	-8.268	-8.412	-1.74	0.048*
QQQ	2023–2024	-7.953	-7.979	-0.33	0.181
QQQ	2025	-7.765	-7.889	-1.60	0.086
GLD	2023–2024	-8.435	-8.402	+0.39	0.001*
GLD	2025	-7.460	-7.345	+1.54	0.070
TLT	2023–2024	-8.150	-8.134	+0.20	0.238
TLT	2025	-8.826	-8.855	-0.33	0.133
EEM	2023–2024	-8.239	-8.240	-0.01	0.949
EEM	2025	-7.896	-8.010	-1.44	0.133
BTC	2023–2024	-6.322	-6.326	-0.06	0.848

Table 4: VaR 1% Attribution Analysis: SPY (2020–2025, 1508 days)

Configuration	Violations	Rate	Improvement
Normal	33	2.2%	—
Student- t (df=5)	18	1.2%	−45.5%
+ Adaptive threshold	14	0.9%	−22.2%
+ Jump augmentation	14	0.9%	0.0%
<i>Target</i>	<i>15</i>	<i>1.0%</i>	

Notes: Violation counts were computed from the frozen 2025-Q4 canonical replication vintage. In the canonical reconstruction, the baseline “Normal” row uses symmetric GARCH(1,1), the Student- t row applies fixed $df = 5$ with scale correction $\sqrt{(df - 2)/df}$, and the adaptive row uses a 20-day rolling maximum of $\hat{\sigma}_t$.

Table 5: VaR Backtest Panel: Joint Pass Rates by Distributional Method and Asset

Method	SPY	QQQ	GLD	TLT	EEM	BTC	IWM	Pass Rate [†]
Skewed- t	✓	✓	✗	✓	✓	✗	✓	90.5% (19/21)
FHS	✓	✓	✓	✗	✓	✓	✓	76.2% (16/21)
CF-VaR	✗	✓	✗	✓	✓	✗	✓	76.2% (16/21)
Student- t (5)	✗	✗	✓	✓	✗	✓	✓	76.2% (16/21)
Normal	✓	✗	✗	✗	✓	✓	✓	57.1% (12/21)

Notes: Each cell summarizes 3 α levels (1%, 2.5%, 5%) $\times$ 3 tests (Kupiec, Christoffersen, DQ) = 9 sub-tests. ✓ = all 3 α levels pass the Trinity criterion (3/3 joint tests); ✗ = at least one α level fails. Total cells: 7 assets $\times$ 3 α $\times$ 5 methods = 105. [†]Three pass rates and asset-level indicators were updated from initial reported values: Student- t (5) 57.1% $\rightarrow$ 76.2% (12/21 $\rightarrow$ 16/21); Skewed- t 76.2% $\rightarrow$ 90.5% (16/21 $\rightarrow$ 19/21); CF-VaR 66.7% $\rightarrow$ 76.2% (14/21 $\rightarrow$ 16/21). Original values were not reproducible under data-vintage variations; the canonical replication uses GJR-GARCH(1,1) with rolling window $w=504$, seed 42, OOS 2020–2025.

Table 6: Volatility Targeting: Cross-Asset Performance

Asset	BH Sharpe	VT Sharpe	Δ Sharpe	BH MaxDD	VT MaxDD
SPY	0.82	0.85	+0.03	−33.7%	−14.8%
GLD	1.56	1.71	+0.15	−25.1%	−13.4%
TLT	0.02	0.33	+0.31	−43.8%	−30.7%
EEM	0.42	0.45	+0.03	−38.2%	−21.5%
BTC	0.43	0.60	+0.17	−76.6%	−21.3%

Notes: Each asset uses its native evaluation window per the analysis in Section C. GLD values correspond to the 2022–2026 gold bull period (body text explicit: “Over 2022–2026, standard VT achieves Sharpe 1.71 versus ...buy-and-hold’s 1.56”). SPY uses 2014–2026 (~12 years; confirmed by BH Sharpe fingerprint match). TLT and EEM use approximately 2015–2026. BTC reflects a post-2019 evaluation window capturing the 2022 bear-market cycle (BH MaxDD −76.6%). Δ Sharpe = VT Sharpe − BH Sharpe. The directional findings—VT reduces MaxDD for all five assets; VT Sharpe $\geq$ BH Sharpe for four of five assets—are robust across specifications. Values reflect the frozen replication vintage; see REPLICATION.md for details on the pinned dataset and per-asset window justification.

References

- Araya, H. T., Aduda, J., & Berhane, T. (2024). A hybrid GARCH and deep learning method for volatility prediction. *Journal of Applied Mathematics*, 2024, 6305525. <https://doi.org/10.1155/2024/6305525>
- Baur, D. G., & Lucey, B. M. (2010). Is gold a hedge or a safe haven? An analysis of stocks, bonds and gold. *Financial Review*, 45(2), 217–229. <https://doi.org/10.1111/j.1540-6288.2010.00244.x>
- Baur, D. G., & McDermott, T. K. (2010). Is gold a safe haven? International evidence. *Journal of Banking & Finance*, 34(8), 1886–1898. <https://doi.org/10.1016/j.jbankfin.2009.12.008>
- Bali, T. G., Engle, R. F., & Murray, S. (2016). *Empirical Asset Pricing: The Cross Section of Stock Returns*. John Wiley & Sons.
- Batten, J. A., Ciner, C., & Lucey, B. M. (2010). The macroeconomic determinants of volatility in precious metals markets. *Resources Policy*, 35(2), 65–71. <https://doi.org/10.1016/j.resourpol.2009.12.002>
- Basel Committee on Banking Supervision. (2006). *International Convergence of Capital Measurement and Capital Standards: A Revised Framework*. Bank for International Settlements.
- Basel Committee on Banking Supervision. (2019). *Minimum Capital Requirements for Market Risk*. Bank for International Settlements.

- Black, F. (1976). Studies of stock price volatility changes. *Proceedings of the Business and Economics Statistics Section, ASA*, 177–181.
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1)
- Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *Review of Economics and Statistics*, 69(3), 542–547. <https://doi.org/10.2307/1925546>
- Bozovic, M. (2024). VIX-managed portfolios. *International Review of Financial Analysis*, 95, 103353. <https://doi.org/10.1016/j.irfa.2024.103353>
- Bucci, A. (2020). Realized volatility forecasting with neural networks. *Journal of Financial Econometrics*, 18(3), 502–531. <https://doi.org/10.1093/jjfinec/nbaa008>
- Cederburg, S., O’Doherty, M. S., Wang, F., & Yan, X. S. (2020). On the performance of volatility-managed portfolios. *Journal of Financial Economics*, 138(1), 95–117. <https://doi.org/10.1016/j.jfineco.2020.04.015>
- Chang, M.-S., Kung, C.-C., Chen, M.-W., & Tian, Y. (2021). Volatility regime, inverted asymmetry, contagion, and flights in the gold market. *Pacific-Basin Finance Journal*, 67, 101522. <https://doi.org/10.1016/j.pacfin.2021.101522>
- Chevallier, J., & Ielpo, F. (2017). Investigating the leverage effect in commodity markets with a recursive estimation approach. *Research in International Business and Finance*, 39, 763–778. <https://doi.org/10.1016/j.ribaf.2014.09.010>
- Engle, R. F., Ghysels, E., & Sohn, B. (2013). Stock market volatility and macroeconomic fundamentals. *Review of Economics and Statistics*, 95(3), 776–797. https://doi.org/10.1162/REST_a_00300
- Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007. <https://doi.org/10.2307/1912773>
- Engle, R. F., & Gallo, G. M. (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics*, 131(1–2), 3–27. <https://doi.org/10.1016/j.jeconom.2005.01.018>
- Acerbi, C., & Szekely, B. (2014). Back-testing expected shortfall. *Risk*, 27(11), 76–81.
- Bayer, S., & Dimitriadis, T. (2022). Regression-based expected shortfall backtesting. *Journal of Financial Econometrics*, 20(3), 437–471. <https://doi.org/10.1093/jjfinec/nbaa013>
- Fissler, T., & Ziegel, J. F. (2016). Higher order elicibility and Osband’s principle. *Annals of Statistics*, 44(4), 1680–1707. <https://doi.org/10.1214/16-AOS1439>
- Patton, A. J., & Sheppard, K. (2015). Good volatility, bad volatility: Signed jumps and the persistence of volatility. *Review of Economics and Statistics*, 97(3), 683–697. https://doi.org/10.1162/REST_a_00503
- Christoffersen, P. F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862. <https://doi.org/10.2307/2527341>

- Christie, A. A. (1982). The stochastic behavior of common stock variances: Value, leverage and interest rate effects. *Journal of Financial Economics*, 10(4), 407–432. [https://doi.org/10.1016/0304-405X\(82\)90018-6](https://doi.org/10.1016/0304-405X(82)90018-6)
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263. <https://doi.org/10.1080/07350015.1995.10524599>
- DeMiguel, V., Martin-Utrera, A., & Uppal, R. (2024). A multifactor perspective on volatility-managed portfolios. *Journal of Finance*, 79(6), 3859–3891. <https://doi.org/10.1111/jofi.13395>
- Engle, R. F., & Siriwardane, E. N. (2018). Structural GARCH: The volatility-leverage connection. *Review of Financial Studies*, 31(2), 449–492. <https://doi.org/10.1093/rfs/hhx099>
- Engle, R. F. (2004). Risk and volatility: Econometric models and financial practice. *American Economic Review*, 94(3), 405–420. <https://doi.org/10.1257/0002828041464597>
- Fleming, J., Kirby, C., & Ostdiek, B. (2001). The economic value of volatility timing. *Journal of Finance*, 56(1), 329–352. <https://doi.org/10.1111/0022-1082.00327>
- Fleming, J., Kirby, C., & Ostdiek, B. (2003). The economic value of volatility timing using realized volatility. *Journal of Financial Economics*, 67(3), 473–509. [https://doi.org/10.1016/S0304-405X\(02\)00259-3](https://doi.org/10.1016/S0304-405X(02)00259-3)
- Francq, C., & Zakoïan, J.-M. (2004). Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli*, 10(4), 605–637. <https://doi.org/10.3150/bj/1093265632>
- Glosten, L. R., Jagannathan, R., & Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48(5), 1779–1801. <https://doi.org/10.1111/j.1540-6261.1993.tb05128.x>
- Hansen, B. E. (1994). Autoregressive conditional density estimation. *International Economic Review*, 35(3), 705–730. <https://doi.org/10.2307/2527081>
- Hansen, P. R., & Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *Journal of Applied Econometrics*, 20(7), 873–889. <https://doi.org/10.1002/jae.800>
- Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2), 453–497. <https://doi.org/10.3982/ECTA5771>
- Hansen, P. R., Huang, Z., & Shek, H. H. (2012). Realized GARCH: A joint model for returns and realized measures of volatility. *Journal of Applied Econometrics*, 27(6), 877–906. <https://doi.org/10.1002/jae.1234>
- Harri, A., & Brorsen, B. W. (2009). The overlapping data problem. *Quantitative and Qualitative Analysis in Social Sciences*, 3(3), 78–115.
- Harvey, C. R., Liu, Y., & Zhu, H. (2016). ...and the cross-section of expected returns. *Review of Financial Studies*, 29(1), 5–68. <https://doi.org/10.1093/rfs/hhv059>
- Harvey, C. R., Hoyle, E., Korgaonkar, R., Rattray, S., Sargaison, M., & Van Hemert, O. (2018). The impact of volatility targeting. *Journal of Portfolio Management*, 45(1), 14–33. <https://doi.org/10.3905/jpm.2018.45.1.014>

- Hood, B., & Raughtigan, C. (2025). Volatility targeting is trendy: How trend following explains alpha in volatility-managed strategies. *Journal of Portfolio Management*, forthcoming. <https://doi.org/10.3905/jpm.2025.1.764>
- Henriksson, R. D., & Merton, R. C. (1981). On market timing and investment performance. II. Statistical procedures for evaluating forecasting skills. *Journal of Business*, 54(4), 513–533. <https://doi.org/10.1086/296144>
- Hwang, S., & Valls Pereira, P. L. (2006). Small sample properties of GARCH estimates and persistence. *European Journal of Finance*, 12(6–7), 473–494. <https://doi.org/10.1080/13518470500039436>
- Kim, T., & Kim, H. Y. (2019). Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data. *PLoS ONE*, 14(2), e0212320. <https://doi.org/10.1371/journal.pone.0212320>
- Kuester, K., Mittnik, S., & Paolella, M. S. (2006). Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, 4(1), 53–89. <https://doi.org/10.1093/jjfinec/nbj002>
- Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models. *Journal of Derivatives*, 3(2), 73–84. <https://doi.org/10.3905/jod.1995.407942>
- Longin, F., & Solnik, B. (2001). Extreme correlation of international equity markets. *Journal of Finance*, 56(2), 649–676. <https://doi.org/10.1111/0022-1082.00340>
- McNeil, A. J., Frey, R., & Embrechts, P. (2015). *Quantitative Risk Management*. Princeton University Press.
- Moreira, A., & Muir, T. (2017). Volatility-managed portfolios. *Journal of Finance*, 72(4), 1611–1644. <https://doi.org/10.1111/jofi.12513>
- Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59(2), 347–370. <https://doi.org/10.2307/2938260>
- Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3), 703–708. <https://doi.org/10.2307/1913610>
- Nelson, R. (2025). Portfolio construction under correlation breakdowns and tail risk. SSRN Working Paper No. 5931154. <https://doi.org/10.2139/ssrn.5931154>
- Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *Journal of Business*, 53(1), 61–65. <https://doi.org/10.1086/296071>
- Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, 160(1), 246–256. <https://doi.org/10.1016/j.jeconom.2010.03.034>
- Sheppard, K. (2023). *arch: ARCH and other financial models in Python* (Version 6.2) [Computer software]. <https://github.com/bashtage/arch>
- Treynor, J. L., & Mazuy, K. K. (1966). Can mutual funds outguess the market? *Harvard Business Review*, 44(4), 131–136.
- Xu, X. (2024). Improving volatility-managed portfolios in real time. *Critical Finance Review*, forthcoming.

Campbell, J. Y., Sunderam, A., & Viceira, L. M. (2017). Inflation bets or deflation hedges? The changing risks of nominal bonds. *Critical Finance Review*, 6(2), 263–301. <https://doi.org/10.1561/104.000000043>