TECHNICAL NOTE

Volatility Surface Calibration and Option Pricing Under Regime, Repair, and Numerical Constraints

Joshua Colmenar

2026-07-12

Contents

1. Abstract and Executive Findings

1.1 Abstract

This report examines the practical problem of converting heterogeneous listed-option quotes into a governed volatility surface and a style-correct pricing book. The calibration problem is not treated as an isolated nonlinear regression. Carry construction, contract routing, objective geometry, static-arbitrage repair, American exercise, numerical convergence, and benchmark design all affect the result. The project therefore evaluates the pipeline as a sequence of conditional numerical decisions rather than searching for one universally superior surface formula.

The canonical European path uses expiry-level put-call parity, forward log-moneyness, raw SVI slices, a Jacobian-consistent total-variance objective, regular-grid evaluation, and a constrained post-fit repair. European contracts are routed through five explicit settlement and expiry families; a pooled label is retained for reporting only. The canonical American path retains Black implied-volatility normalization for calibration but prices exercise explicitly with a Cox-Ross- Rubinstein tree. A matched European tree value provides a numerical early-exercise- premium decomposition. Leisen-Reimer and finite-difference engines are maintained as independent numerical checks rather than daily selectors.

The main surface failure was a tail problem concentrated in nonstandard weekly index contexts, not a general failure of average fit. Stronger direct-price and regularization formulations frequently improved an in-sample objective while worsening diagnostic feasibility or adverse-date repricing. A weak, guarded, price-local SVI refinement was therefore promoted only for eligible `weekly_pm_nonstandard` slices. In the frozen August panel it changed 7 of 24 asset-date slices. Mean repaired RMSE fell from 7.4179 to 3.8995 on the liquidity basket and from 8.6654 to 3.9681 on the representative basket. A de-duplicated full-supported-chain companion fell from 6.3036 to 3.1700 with no loss among changed contexts. A November control month showed smaller improvements of approximately 2.6% and 2.0%; that is a limited transfer check rather than broad temporal validation.

Alternative formulations remained useful without becoming canonical. Repair-matched eSSVI produced a credible `monthly_pm_standard` candidate but not a global replacement. A smoothed local-window repo branch won 80 of 202 exact-support contexts, yet its aggregate repaired-grid RMSE was materially worse than parity in both evaluation baskets. American-aware implied-volatility normalizations changed some local results but created no new promotion context across the available AAPL, SPY, and INTC panels. These outcomes are retained because they constrain future work more usefully than a list of successful optimizations.

The closing panel contains 48 asset-date contexts and 2,400 valid prices. Aggregate raw-price RMSE is 3.5122 and MAE is 1.4620, but those pooled values are scale dependent and are reported with per-asset results. The 1,200 matched American pairs have mean numerical early-exercise premium 0.004264 and no negative-premium or intrinsic-bound failure. The strict release gate nevertheless fails: seven quotes sit exactly at the 40% relative-spread limit. The system is consequently described as a functional research and controlled batch-pricing system, not an unattended trading or risk platform.

1.2 Executive Findings

1. **The canonical method is strong for identifiable reasons.** Parity imposes an economic cross-option restriction; forward coordinates remove a large deterministic term-structure component; the Jacobian weight aligns total-variance residuals with quote-price uncertainty; and grid repair removes selected discrete static inconsistencies. These are legitimate advantages, not evidence that every comparison was initially fair.
2. **Comparison design mattered.** Early studies sometimes compared repaired canonical grids with unrepaired challengers, pooled unlike expiries, or scored an export basket without a full-chain companion. Those designs were corrected. Repaired-versus-repaired symmetry, atomic family routing, common contexts, exact support, and explicit denominator disclosure are now required.
3. **Average RMSE was an inadequate promotion criterion.** Several challengers improved mean fit while creating severe family-date losses. Tail controls, diagnostics, and no-harm guards became first-class acceptance conditions.
4. **The weekly improvement is narrow by design.** The promoted refinement is an eligible-slice policy inside raw SVI, not a new global surface family. It changes at most one qualifying slice per family context and must pass variance, price, parameter, butterfly, calendar-neighbor, and support guards.
5. **Carry and basket reuse were audited, not assumed harmless.** The carry audit covers 357 expiries. Its mean interleaved-strike cross-validation overfit ratio is 1.043, the mean absolute log-forward shift is 0.0337 basis points, and no aggregate selection or keep/drop flip occurs. The basket audit shows the export basket is only about 5.98% of full-supported rows on average, so both views remain necessary. These results reduce specific endogeneity concerns but do not prove exogeneity on unobserved history.
6. **American pricing is explicit downstream even though normalization remains approximate.** Black inversion remains the calibration baseline because American- aware alternatives did not generalize. CRR, Leisen-Reimer, and CN-PSOR preserve exercise semantics in pricing and numerical validation.
7. **The project is finished as a bounded research system.** It installs, tests, and runs on synthetic input without proprietary access. It does not claim live-data operation, independent long-history validation, transaction-cost-aware trading profitability, or service-level reliability.

2. Problem Definition and Scope

2.1 The object being estimated

For each observation date and underlying, the input is a cross-section of call and put quotes spanning strikes, expiries, settlement conventions, and exercise styles. The immediate statistical object is a total-implied-variance surface. The operational object is broader: a pricing state that carries forward, discount, surface, interpolation, repair, exercise, numerical-engine, and quality metadata without silently changing conventions between stages.

A surface can fit quoted implied volatilities and still be unsuitable for pricing. Three mechanisms are central. First, implied volatility is a nonlinear coordinate of price. Equal errors in total variance do not have equal dollar impact across strikes. Second, independently calibrated slices can cross in maturity or imply locally nonconvex strike prices after interpolation. Third, American exercise makes the normalization and pricing operators model dependent. The project treats these as separate questions rather than asking one optimizer to absorb all of them.

The canonical calculation is therefore conditional on a declared state `S_d = {Q_d, F_d, D_d(T), forward_d(T), Theta_d, G_d, Pi_d, E_d}`, where Q_d is the normalized quote set, F_d the contract-family routing, D_d and F_d the carry state, Θ_d the fitted slice parameters, G_d the evaluated grid, Π_d the repair policy, and E_d the exercise and numerical-engine policy. This notation is a project organization device, not a claim that the state is sufficient for the true market data-generating process.

2.2 Research questions

The work addresses five linked questions.

1. Can European and American contracts be routed without pooling economically different settlement and expiry mechanisms?
2. Can a stable SVI-based surface preserve quote information while controlling discrete calendar and butterfly defects?
3. Does aligning the calibration objective with local price sensitivity improve pricing without introducing tail instability?
4. Can American exercise and its premium be priced and diagnosed independently of the normalization convention used to create the surface?
5. Can candidate methods be compared without repair asymmetry, support changes, basket endogeneity, or daily best-model look-ahead?

The fifth question materially changed the answers to the first four. An optimizer's reported success was never accepted as a promotion result. A candidate had to be compared on common contexts and support, under matched downstream operators, with tail and diagnostic checks. Historical gates were treated as project policies with provenance, not as universal statistical constants.

2.3 System boundary

The frozen public package supports quote normalization, explicit contract classification, parity and auxiliary carry, implied-volatility roots, SVI fitting, grid repair, European and American pricing, early-exercise decomposition, quality classification, benchmark generation, and controlled research comparators. It also contains a C++17 pricing library with Black, CRR, Leisen-Reimer, fully implicit, Crank-Nicolson, and CN-PSOR engines.

The package does not include raw market chains, vendor identifiers, credentials, or private data-access code. Its public empirical evidence consists of sanitized context-level or sufficient-statistic tables. Synthetic examples exercise the full public API. This boundary is deliberate: reproducibility means that

every published number can be recomputed from public derived data, not that licensed source rows are redistributed.

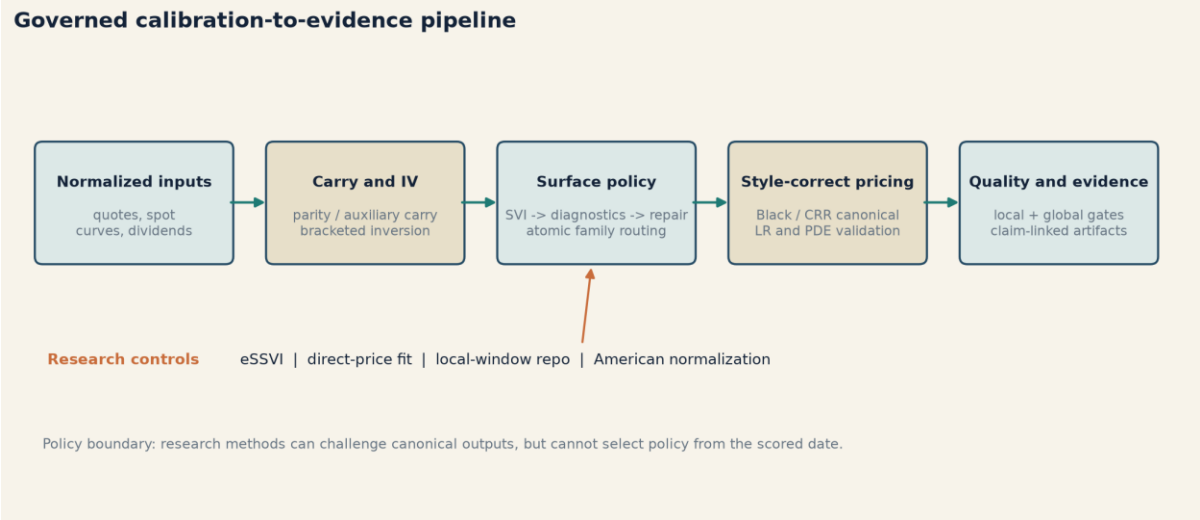

Figure 1. Governed calibration-to-evidence pipeline. The research branch can challenge carry, normalization, surface, and objective choices, but it cannot enter canonical output without explicit promotion. Arrows describe calculation flow, not causal identification.

2.4 Appropriate interpretation

The project is functional in the narrow engineering sense: its inputs, state transitions, outputs, failure modes, and tests are explicit. It is also functional in the research sense: promoted and rejected formulations can be reconstructed from a frozen registry. That does not make it production complete. There is no live market- data entitlement, unattended scheduler, persistent service, operational monitoring, hedge execution, transaction-cost model, or independently sampled multi-year test.

The intended interpretation is therefore a controlled pricing research system and a reference implementation. Its strongest contribution is not a single error number. It is the set of quantitative distinctions needed to avoid a misleading result: European versus American exercise, raw versus repaired grids, atomic families versus reporting pools, export baskets versus full chains, local wins versus global policy, and numerical reference values versus ground truth.

3. Data Conventions, Carry, and Forward Coordinates

3.1 Available panels

The private archive at the freeze contains 238 root-level option files across seven assets, 559 auxiliary CSV files, and three root work files that satisfy neither definition. Coverage is irregular. The dates are available observation windows, not a probability sample of market regimes. The public evidence preserves file counts, date ranges, and context-level aggregates but no quote-level prices, strikes, stable contract identifiers, or private paths.

Table 1. Available raw option coverage at the freeze.

Asset	Raw files	First date	Last date	Distinct dates	Market role
AAPL	5	2024-09-16	2024-09-20	5	American equity
INTC	55	2024-07-01	2025-08-29	55	American equity
NDX	73	2024-07-01	2025-08-29	73	European index
NVDA	8	2025-08-20	2025-08-29	8	American equity
RUT	31	2024-11-04	2025-08-29	31	European index
SPX	53	2024-07-01	2025-08-29	53	European index
SPY	13	2024-09-16	2025-08-29	13	American ETF

Raw-dollar error comparisons are meaningful within an asset and useful for an operational pricing book, but they are not scale free. A one-dollar error on SPX and a one-dollar error on INTC do not represent equivalent relative quality. The report therefore uses pooled values only as accounting summaries and keeps per-asset and family panels visible.

3.2 European parity carry

For a fixed European expiry, matched call and put prices at common strikes satisfy put-call parity. The implementation estimates the relation

$$C(K,T) - P(K,T) = \alpha(T) + \beta(T)K. \qquad (1)$$

Under the forward-price convention used in the package,

$$D(T) = -\beta(T), \qquad F(T) = \frac{\alpha(T)}{D(T)}, \qquad r(T) = -\frac{\log D(T)}{T}. \qquad (2)$$

The sign matters. Since the coefficient on strike is -D(T), a positive fitted slope is not clipped into a plausible discount factor; it is a failed carry state. The public unit audit recovers both discount and forward from an exact synthetic parity system to machine precision.

Parity has two useful properties in this setting. It uses the cross-option economic restriction directly, and it produces a local expiry-level forward rather than requiring a globally parameterized dividend yield. It also creates an endogeneity risk: the same chain can contribute to carry estimation, quote selection, surface fit, and scoring. Section 8 evaluates that risk rather than treating parity as an external input.

3.3 Rate curves and cash dividends

American equity and ETF panels use saved auxiliary inputs consisting of a zero-rate curve and forecast cash dividends. Zero rates are interpolated by a shape-preserving piecewise cubic Hermite method. Let cash dividend D_i occur at t_i $\leq$ T. The project coordinate is

$$\mathrm{PV}_{\mathrm{div}}(T) = \sum_{t_i \le T} D(t_i) D_i, \qquad F(T) = \frac{S - \mathrm{PV}_{\mathrm{div}}(T)}{D(T)}, \qquad q(T) = r(T) - \frac{\log(F(T)/S)}{T}. \qquad (3)$$

This is a forward construction, not a claim that discrete dividends are equivalent to a continuous yield for every American exercise problem. The downstream tree and finite-difference interfaces currently receive the resulting equivalent yield. That approximation is one reason the American normalization and

ex-dividend-sensitive research remain explicitly limited. Known cash dividends can make pre-dividend call exercise optimal; Whaley's correction to earlier closed-form work is a useful warning against casual dividend substitutions.

The public package exposes vendor-neutral curve and dividend schemas but cannot refresh the saved private auxiliary data. Synthetic fixtures demonstrate the same calculation without implying that the historical curve is redistributable.

3.4 Forward coordinates

Once carry is fixed, the surface uses log-forward moneyness and total variance:

$$k = \log\left(\frac{K}{F(T)}\right), \qquad w(k,T) = \sigma(k,T)^2 T. \qquad (4)$$

These coordinates remove deterministic discount and carry effects from the strike axis and put maturities on a common variance scale. They do not remove carry risk. An error in $F(T)$ shifts every strike in the slice and changes both the observed smile and the Jacobian weights. Carry must therefore be part of the benchmark state, not a preprocessing detail that disappears from later attribution.

3.5 Time, price, and option conventions

Maturity is measured in year fractions under the configured 365.25-day basis. Option type is represented as $\omega = +1$ for a call and -1 for a put. Quote midpoints are used for the simplified research comparisons, while half-spread enters weighting and quality classification. Invalid ordering, nonpositive coordinates, out-of-bound prices, insufficient parity pairs, failed roots, or invalid tree probabilities produce typed failures. They are not converted into plausible values by clipping.

This failure policy is operationally important. A calibration dataset composed only of successful inversions is already selected. Counts of rejected roots, excluded support, and invalid prices must accompany fit statistics so a lower RMSE cannot be created by silently shrinking the evaluation domain.

4. Contract-Family and Settlement Taxonomy

4.1 Why tenor pooling was rejected

An option expiry label does not uniquely determine its economic convention. Standard monthly, explicit month-end, quarter-end, and nonstandard weekly contracts can differ in root, settlement time, liquidity, forward sensitivity, and quote density. Pooling them into a generic weekly or mixed bucket makes a calibration score easier to compute but harder to interpret. A favorable average can conceal a tail concentrated in one convention; a fitted parameter can be forced to absorb a settlement difference that should have been handled by routing.

The final European classifier uses three explicit axes:

1. settlement: `am` or `pm`;
2. root class: `standard_root` or `weekly_root`;

3. expiry archetype: `standard_monthly`, `month_end_explicit`, `quarter_end_explicit`, or `weekly_nonstandard`.

The resulting atomic calibration families are shown in Table 2. `pooled` is a union for reporting and never a calibration target. American equity and ETF contracts retain separate style-aware routing rather than being forced through the European index taxonomy.

Table 2. European index calibration families.

Family	Settlement	Root class	Expiry archetype	Canonical surface
`monthly_am_standard`	AM	standard	standard monthly	repaired grid
`monthly_pm_standard`	PM	weekly	standard monthly	repaired grid
`weekly_pm_nonstandard`	PM	weekly	nonstandard weekly	repaired grid plus guarded local refinement
`eom_pm`	PM	weekly	explicit month-end	repaired grid
`quarter_end_pm`	PM	weekly	explicit quarter-end	repaired grid
`pooled`	mixed	mixed	reporting union	no calibration

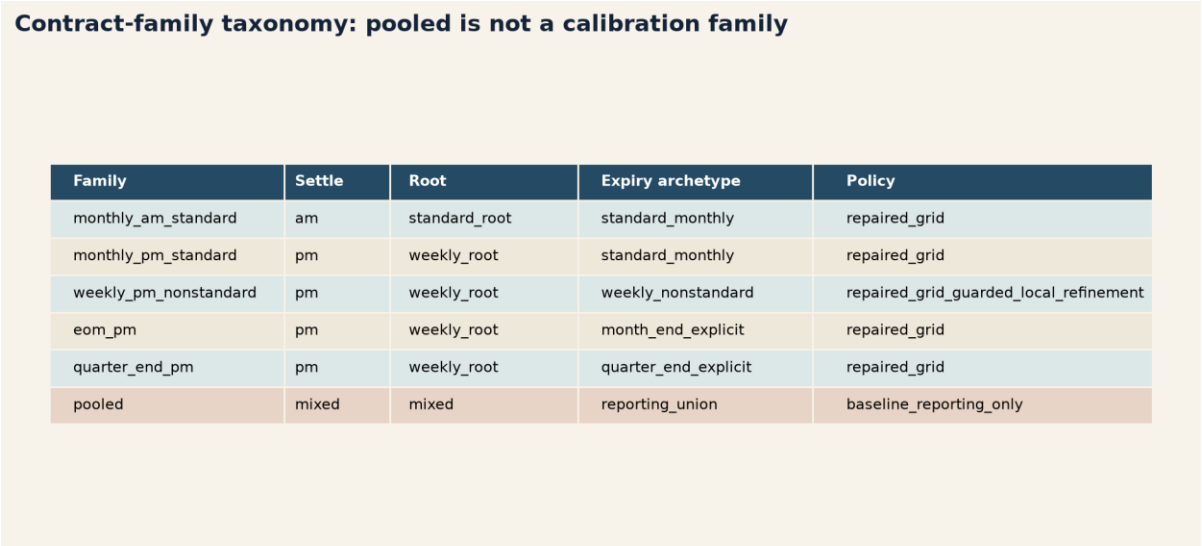

Figure 2. Contract-family taxonomy. Classification is an input to calibration, not a label applied after scoring.

4.2 Classification rules

Standard monthly expiry is identified by its calendar rule and settlement/root combination. Explicit month-end contracts require the month-end indicator rather than being inferred solely from proximity to the last business day. Quarter-end is a subclass of explicit month-end and is separated because March, June, September, and December flows can differ. A nonstandard weekly is PM settled, uses the weekly root, is not a standard monthly date, and is not explicitly month-end.

The implementation deliberately uses a weekday month-end test rather than embedding an exchange holiday calendar in the core classifier. That is a documented limitation: a vendor-grade production classifier should use exchange calendars and contract master data. In the frozen panels, the explicit expiry indicator and root fields provide the main separation.

4.3 Statistical role of the taxonomy

The classifier does more than improve reporting. It defines the unit over which surface policy, repair, paired comparison, and promotion gates are applied. A candidate is not allowed to offset a severe weekly loss with a monthly gain and call the pooled result acceptable. Conversely, a useful method can remain a local family candidate even when it is not suitable globally. Repair-matched eSSVI illustrates this distinction: its `monthly_pm_standard` result survives family-level gates, while the global replacement does not.

Atomic routing also clarifies denominator choice. A family-date context is the paired unit for most surface comparisons. Quote-level errors are summarized within that context before cross-context means and bootstrap intervals are computed. This avoids allowing a single dense chain to dominate merely because it has more strikes.

4.4 Policy implication

The taxonomy is promoted infrastructure. It is not a learned model and is not selected daily. Any future family addition requires an explicit definition, a market convention rationale, configuration wiring, benchmark support, and a decision about whether it is a calibration family or a reporting union. Catch-all buckets are not permitted to become calibration families by convenience.

5. Implied-Volatility Inversion and SVI Calibration

5.1 Black forward valuation

For $\omega=+1$ for calls and -1 for puts, the discounted Black forward value is

$$d_1 = \frac{\log(F/K) + \frac{1}{2}\sigma^2 T}{\sigma\sqrt{T}}, \qquad d_2 = d_1 - \sigma\sqrt{T}, \qquad V = D\,\omega[FN(\omega d_1) - KN(\omega d_2)]. \qquad (5)$$

European lower bounds are $\mathrm{Dmax}(\mathrm{F\text{-}K},0)$ for calls and $\mathrm{Dmax}(\mathrm{K\text{-}F},0)$ for puts; the common upper bounds are DF and DK. American lower bounds use immediate spot exercise. These checks occur before root finding. A quote outside the style-correct interval is a data or convention failure, not an extreme volatility.

Black's forward form is used for European normalization and pricing. For American quotes it remains a calibration coordinate, not an American valuation claim. The downstream pricing book preserves the exercise style separately.

5.2 Bracketed implied volatility

Implied volatility solves

$$f(\sigma) = V_{\text{model}}(\sigma) - V_{\text{market}} = 0. \qquad (6)$$

The canonical solver uses Brent's bracketed method with initial volatility interval $[10^{-6},3]$, controlled upper expansion to 5 and 8, absolute root tolerance 10^{-8}, and at most 80 iterations. The initial endpoints are not interpreted as economic volatility bounds. They are numerical brackets, and expansion is explicit. If a sign change cannot be established, the result records a bracket failure. A synthetic Black round trip reproduces the input volatility to approximately 3.4×10^{-10}.

American tree inversion uses the same bracket discipline but first searches for the lowest volatility at which the discrete risk-neutral probability is admissible. It does not reject a valid market root merely because an arbitrarily small starting volatility makes one tree configuration invalid. The tree itself never clips an inadmissible probability into $[0,1]$.

5.3 Raw SVI

Each accepted expiry slice is represented in raw SVI total variance:

$$w(k) = a + b\left[\rho(k-m) + \sqrt{(k-m)^2 + \sigma_{\text{svi}}^2}\right]. \qquad (7)$$

The parameter bounds enforce $b>0$, $|\rho|<1$, and $\sigma\_svi>0$, together with finite bounds on level and location. These restrictions are necessary for a usable fit but do not by themselves guarantee global absence of butterfly arbitrage.

For a smooth total-variance smile, the density-related SVI diagnostic is

$$g(k) = \left(1 - \frac{kw'(k)}{2w(k)}\right)^2 - \frac{w'(k)^2}{4}\left(\frac{1}{w(k)} + \frac{1}{4}\right) + \frac{w''(k)}{2}. \qquad (8)$$

Nonnegative g, positive total variance, and appropriate wing behavior are linked to butterfly-arbitrage control. The implementation evaluates analytic raw-SVI derivatives on a dense support grid and also maintains a finite-difference version for repaired grids. It calls these checks diagnostics or proxies where the discrete operator does not establish a continuous global theorem.

5.4 Optimization and acceptance

Raw SVI is estimated with deterministic multistart L-BFGS-B. Each start is bounded, and the best successful objective is retained. Optimizer success is only the first acceptance condition. The result must also have finite parameters, positive variance on dense support, sufficient quote count and width, bounded parameter fragility, and configured butterfly diagnostics.

This separation prevents a common numerical mistake: treating convergence of a nonlinear optimizer as evidence that the fitted object is economically or operationally admissible. A local minimum can have a low residual and still sit near a parameter boundary, imply a fragile wing, or fail after interpolation.

5.5 American normalization limitation

The canonical surface path still converts American quote mids through Black implied-volatility inversion. That choice is internally consistent as a coordinate but omits the early-exercise premium from the inversion. CRR American IV and a premium-adjusted Black target were implemented and tested. Their effects were localized and did not create a new promotion context across the available AAPL, SPY, and

INTC panels. The baseline was retained, and the limitation is carried into the pricing report rather than hidden by relabeling the normalization as American.

6. Jacobian-Consistent Weighting and Price-Space Alignment

6.1 The objective mismatch

SVI is fitted in total-variance space, while the downstream score is often a dollar price error. A uniform squared error in w does not correspond to a uniform price error. Near-the-money contracts with high vega can be much more price sensitive than far-wing contracts, and a fixed bid-ask width has different informational meaning at different price levels.

Let $w=\sigma^2T$. By the chain rule,

$$\frac{\partial V}{\partial w} = \frac{\partial V}{\partial \sigma}\frac{\partial \sigma}{\partial w} = \frac{\text{Vega}}{2\sigma T}. \qquad (9)$$

If h_i is the quote half-spread and e_i=w(k_i;θ)-w_i, the local standardized price residual is approximately

$$\frac{V(w_i + e_i) - V(w_i)}{h_i} \approx \frac{\text{Vega}_i}{2\sigma_i T_i h_i} e_i.$$

The resulting Jacobian-consistent total-variance weight is

$$\lambda_i \propto \left(\frac{\text{Vega}_i}{2\sigma_i T_i h_i}\right)^2. \qquad (10)$$

The implementation floors unstable denominators and normalizes weights within a slice. A finite-difference audit confirms the price-per-total-variance derivative to approximately 2.9×10^{-10} in the packaged synthetic calculation.

6.2 Calibration objective

The controlled SVI objective is

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \lambda_i [w(k_i;\theta) - w_i]^2, \qquad (11)$$

where Θ is the bounded raw-SVI parameter set. This is a first-order price-space alignment, not direct price fitting. It preserves the smooth and cheap total-variance representation while making the residual metric more consistent with quoted price uncertainty.

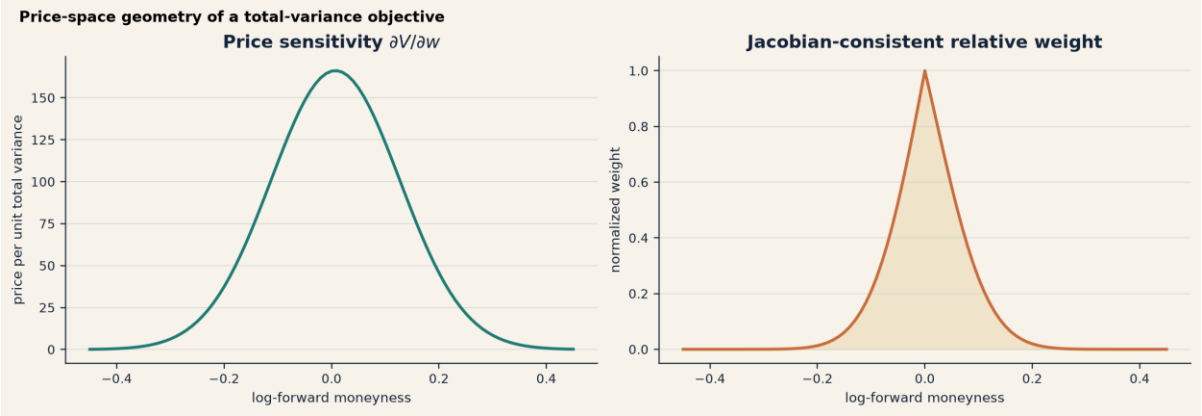

Figure 3. Price-space geometry of total-variance fitting. The synthetic example shows how Black vega, maturity, volatility, and spread transform an equal total-variance error into highly unequal price significance. It is an explanatory calculation, not an empirical distribution of weights.

6.3 What the Jacobian does and does not solve

The weight improves local metric alignment. It does not make the SVI model linear, make the first-order approximation exact for large residuals, or ensure that the resulting grid is free of static defects. It can also concentrate influence in a small region if vega is high and spreads are narrow. Parameter and influence diagnostics remain necessary.

The weight became the controlled baseline because its rationale is identifiable and its later comparisons were more coherent than ad hoc residual weights. This does not mean every historical gain should be attributed to the Jacobian. Taxonomy fixes, support handling, repair corrections, and changed benchmark governance occurred over the same research program. The experiment registry separates those interventions.

6.4 Why direct price fitting was still tested

The Jacobian is a local approximation. Direct price fitting removes that approximation by evaluating Black prices inside the optimizer. It was therefore a logical challenger. However, it creates a harder nonconvex problem and can trade a small price-residual gain for unstable SVI parameters or shape violations. The project's direct-price branch used spread-scaled Huber residuals, parameter proximity, positive-variance penalties, and a butterfly guard. It improved selected quote metrics but failed broad diagnostic or tail promotion requirements. Section 11 reports that result as evidence about the objective-model interaction, not as a failed attempt to minimize prices.

6.5 Quantitative interpretation

The correct interpretation of the canonical objective is conditional: among bounded raw-SVI slices fitted to accepted implied volatilities, it gives more influence to observations whose total-variance error has greater local price significance relative to spread. It is not a likelihood derived from a complete quote-noise model. Half- spread is used as an operational precision scale, and serial, cross-strike, and cross-expiry dependence are not estimated.

This distinction matters for inference. The paired bootstrap in later sections resamples context-level score differences; it does not treat the weighted calibration residuals as independent Gaussian observations. Calibration weights and benchmark uncertainty answer different questions and are kept separate.

7. Static-Arbitrage Diagnostics and Repaired-Grid Construction

7.1 Why repair is separate from calibration

Independent SVI slices can fit their quoted expiries and still produce an evaluated grid with decreasing total variance in maturity or a negative strike-density proxy. The project handles this with a post-fit grid projection. It does not call the projection another calibration because no quote objective is re-estimated. The input is the fitted total-variance grid, the output is a nearby grid satisfying configured discrete conditions, and the displacement is recorded.

This distinction has two consequences. First, repair can improve static consistency while worsening quote repricing. Second, a repaired grid is not proof that the underlying continuous SVI surface is globally arbitrage free. The operator enforces conditions on a declared grid and support domain. The report therefore uses `repaired_grid` as a policy label, not as a theorem.

7.2 Calendar projection

At each fixed log-moneyness column k_j, total variance should not decrease with maturity under the configured calendar diagnostic. The repair computes the isotonic least-squares projection

$$\widehat{\mathbf{w}}_j = \arg\min_{\mathbf{z}} \sum_{\ell=1}^{L} (z_\ell - w_{\ell j})^2 \quad \text{subject to} \quad z_1 \le z_2 \le \cdots \le z_L. \qquad (12)$$

The pool-adjacent-violators algorithm solves this one-dimensional projection deterministically. Calendar repair is applied column by column. It preserves neither raw-SVI parameters nor exact quote fit; it preserves proximity to the evaluated grid under the stated norm.

7.3 Discrete strike diagnostic

For each maturity row, finite-difference estimates of w' and w'' are inserted into the SVI density function. On grid node k_j, the diagnostic is

$$g_j = \left(1 - \frac{k_j \widehat{w}'_j}{2w_j}\right)^2 - \frac{(\widehat{w}'_j)^2}{4}\left(\frac{1}{w_j} + \frac{1}{4}\right) + \frac{\widehat{w}''_j}{2}. \qquad (13)$$

Rows with g_j below tolerance are projected by constrained SLSQP. For candidate row z, the objective is

$$\min_z \ \|z - w\|_2^2 \quad \text{subject to} \quad z_j > 0,\ g_j(z) \ge -\varepsilon_g, \qquad (14)$$

with bounds that prevent nonpositive total variance. Because the calendar and strike constraint sets are coupled, the full operator alternates calendar and row projections. It reports convergence only when the final combined diagnostics pass.

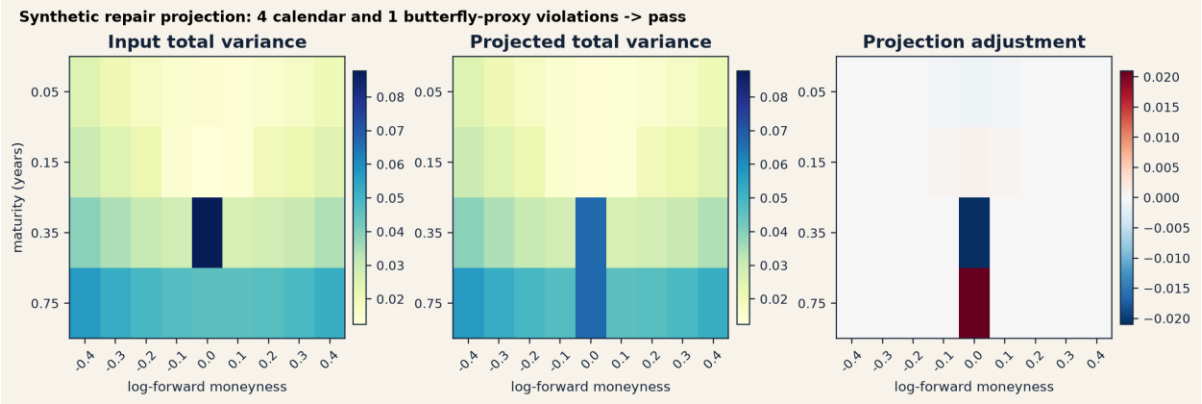

Figure 4. Deterministic repair example. The input contains four calendar violations and one butterfly-proxy violation. The final grid passes the configured discrete checks. The example demonstrates operator behavior and is not a market surface.

7.4 Failure semantics

A failed European repair does not silently fall back to the raw grid under a repaired label. The result carries a failed status, the last diagnostics, iteration count, and maximum displacement. This was a necessary correction. Earlier repair-path and support defects affected monthly and weekly comparisons; they were fixed and rerun before the current reference was established.

The first dedicated no-repair experiment is also excluded. A Windows output-path error contaminated its artifacts, so its numerical comparison is inadmissible. The corrected `norepair2` run is the valid control. Keeping both entries in the experiment registry matters: deleting the invalid run would make later result changes difficult to explain, while using it would violate the project's own artifact-integrity rule.

7.5 Fair comparison rule

Once repair became canonical for European atomic families, comparing a repaired control with an unrepaired challenger became a confounded experiment. A method could lose because of its fit, because it lacked the same projection, or because support changed after repair failure. Later eSSVI, joint-SVI, direct-price, and repo studies therefore use a common repair wrapper where the mathematical output is compatible with that operator.

Matched repair does not guarantee a fair study by itself. The two branches must still share contracts, family routing, pricing engine, carry state unless carry is the treatment, and aggregation rules. The local-window repo study required an additional private exact-contract join because branch-specific failures had changed quote support. Only context sufficient statistics crossed the public boundary.

7.6 Limits of the operator

The repair is deliberately conservative. It does not infer a stochastic-volatility model, estimate local volatility, or guarantee arbitrage-free extrapolation beyond the grid. It can shift values away from quotes, and alternating projections need not solve a single globally convex problem when the discrete g constraints are nonlinear. The operator is accepted because its constraints and displacement are inspectable, its failures are explicit, and its effect can be benchmarked against the raw grid.

8. Benchmark Governance, Representativeness, and Endogeneity Controls

8.1 Governance replaced a fixed leaderboard

The early benchmark was useful as a historical control but was not retained as an unchanging definition of success. Its family labels, weighting, repair assumptions, and gates predated later corrections. Treating those values as exogenous truth would have made the research circular: a candidate would be judged against policies partly created by the incumbent method.

The current governance policy is versioned as `2026-04-13-rigor-v1`. Its gates are engineering and research controls, not natural constants. A change to a gate requires a documented reason, sensitivity analysis, and frozen-panel rerun. Passing a newly chosen threshold is not sufficient justification for the threshold.

8.2 Unit of comparison

Surface candidates are paired by asset, observation date, atomic contract family, and evaluation basket. They must share the pricing convention and, except in a carry experiment, the carry state. Quote-level errors are first aggregated within context. The paired context differences are then summarized by mean, lower-tail quantile, worst adverse value, win rate, and a paired bootstrap interval.

The sign convention is explicit: positive `canonical RMSE minus candidate RMSE` favors the candidate. This prevents a common reporting error in which improvements change sign between tables. Bootstrap intervals describe the observed context sample; they do not repair the irregular historical sampling design.

Table 3. Promotion-gate roles.

Gate	Question	Typical evidence	Failure implication
A: paired mean	Is average paired error lower?	mean delta and paired interval	no broad average advantage
B: tail	Are adverse contexts controlled?	lower quantile and worst loss	candidate can improve mean but remain unsafe
C: diagnostics	Is numerical/surface quality no worse?	warning and failure rates	fit gain is not operationally admissible
D: cross-basket	Does the conclusion survive a second sample rule?	both export baskets and companion chain	result may be selection-specific

The letters are report shorthand for the frozen gate table. They should not be read as a universal statistical test battery. In particular, a family may remain a research candidate after failing global promotion.

8.3 Basket endogeneity

The operational benchmark exported a controlled set of up to 50 quotes per context. That improves comparability and runtime but can create selection dependence. The basket audit recomputed the same fitted surfaces on every contract inside supported coordinates.

The candidate audit contains 96 context-surface-basket rows. The mean export-to-full- supported count ratio is 0.05983. Full-chain values are repeated across the two export-basket labels in that audit table, so a naive 96-row average would double the apparent denominator. Policy comparisons use a separate de-duplicated 24-context full-chain table.

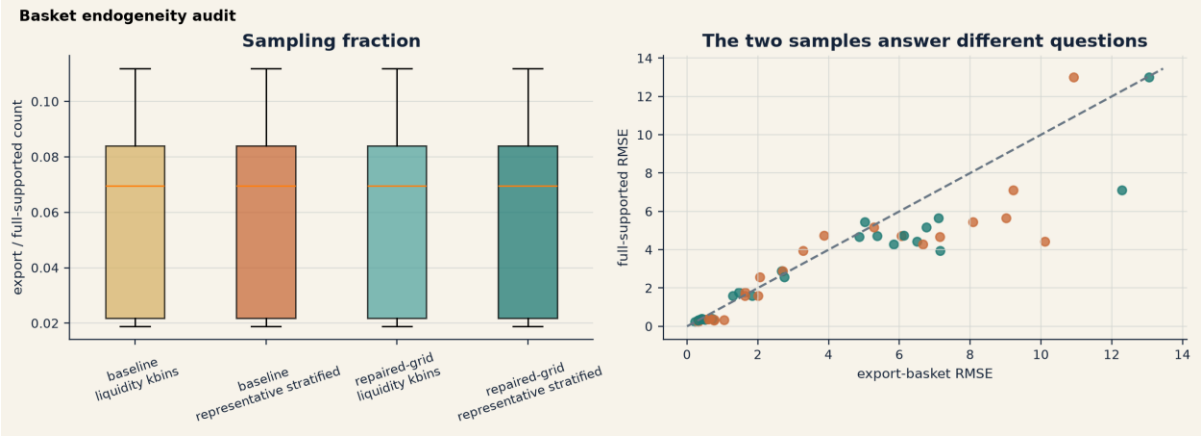

Figure 5. Export-basket and full-supported-chain comparison. The two samples answer different questions; repeated full-chain values in the audit denominator are disclosed rather than treated as independent observations.

The result is not that export selection is harmless. Export and full-chain scores can differ materially. The governance response is to require both: the controlled basket answers whether a fixed operational sample improves, while the companion chain checks whether the surface change merely moved error into unexported strikes.

8.4 Carry endogeneity

European parity uses option prices from the same expiry later used for selection and surface fitting. The carry audit constructs deterministic interleaved-strike folds, refits the parity state without each holdout fold, and records forward perturbation, fit overstatement, selection-state changes, and keep/drop changes.

Across 357 valid expiries, the mean cross-validation overfit ratio is 1.04298 and the mean absolute log-forward shift is 0.03374 basis points. Aggregate selection and keep/drop flip rates are zero under the audit definitions. These are small effects relative to the observed weekly repricing differences. They support the inference that parity reuse is not the primary source of canonical strength on this panel. They do not establish causal exogeneity, independence across strikes, or stability in unobserved market periods.

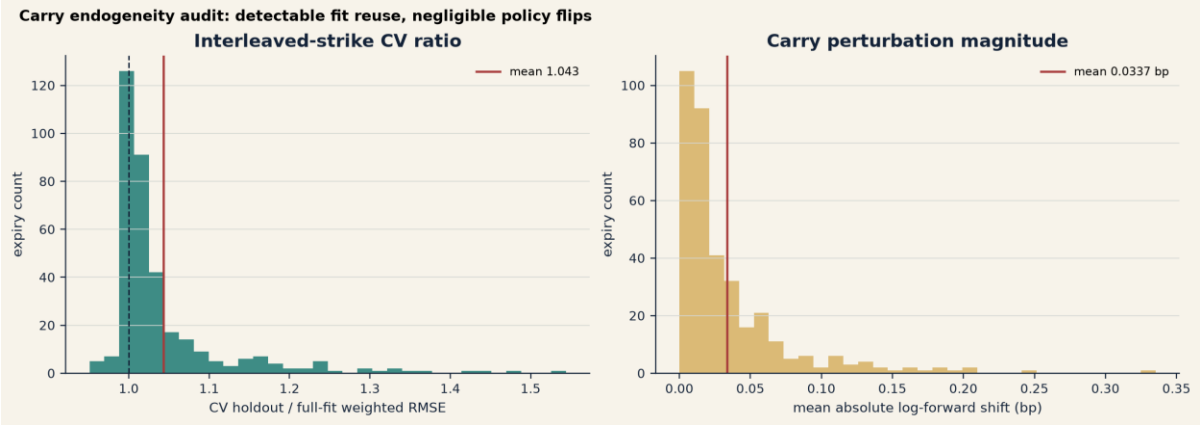

Figure 6. Carry endogeneity audit over 357 valid expiries. Cross-validation is a deterministic interleaved-strike stress, not a causal instrument for carry.

8.5 Look-ahead controls

The final canonical policy does not select the best surface or pricing engine after observing each day's benchmark result. European and American engines are fixed by exercise style. The weekly refinement is governed by predeclared eligibility and no-harm conditions and can change at most one eligible slice in a context. It is not a daily oracle over the candidate library.

The research process did inspect a frozen August panel while designing the policy. That creates model-development dependence even without daily look-ahead. The November panel is a limited transfer check, not a clean large-sample holdout, and the closing August panel overlaps the available research era. The report therefore uses `promoted narrow policy` rather than `out-of-sample proven model`.

8.6 Denominators and scale

Every displayed aggregate identifies whether the denominator is quotes, contexts, expiries, asset-dates, or matched pairs. Raw-dollar RMSE is not pooled across assets to claim economic superiority. Where a pooled value is retained for bundle accounting, per-asset values and a scale caveat accompany it. These conventions were introduced because average fit and unlabeled pools repeatedly concealed the actual failure mechanism.

9. Root-Cause Hypotheses and Controlled Tests

9.1 Research sequence

The project records 32 formal experiments. The sequence is not a linear progression from weak to strong models. It alternates between software corrections, comparison- design corrections, model challengers, targeted diagnoses, and numerical validation. The distinction matters because a repair bug fix and a model promotion are different claims even if both lower an error metric.

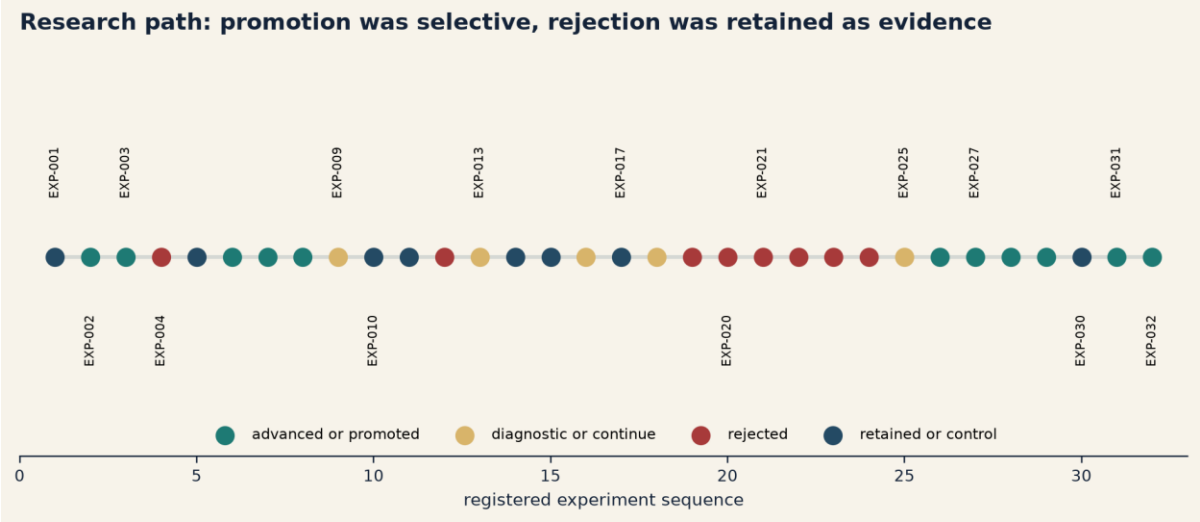

Figure 7. Frozen research decision path. The diagram records decision classes and sequence; spacing is neither elapsed time nor statistical distance.

9.2 Main hypotheses

Table 4. Root-cause hypotheses and dispositions.

Hypothesis	Test	Finding	Disposition
Pooled expiry families dilute a localized tail	atomic settlement/root/expiry routing	family-specific weekly and monthly behavior became visible	classifier promoted
Repair implementation explains adverse monthly behavior	repair-path audit and rerun	a concrete defect was found and fixed	software fix promoted
Weekly support handling explains the remaining tail	supported-coordinate correction and common-support studies	support handling mattered but did not eliminate tail losses	fix promoted; hypothesis incomplete
Canonical strength is mainly carry reuse	interleaved-strike parity refits	forward shifts and selection changes were negligible on audited panels	primary-cause hypothesis not supported
Export basket creates the weekly gain	full-supported-chain repricing	export and full chain differ, but targeted changed contexts retained gains	companion chain required
Boundary contracts drive the adverse weekly dates	boundary squared-error attribution	endpoint contribution was small on the principal target and November control	primary-boundary hypothesis rejected
A more coherent global surface fixes the problem	eSSVI and joint SVI with matched repair	local candidates emerged; no global replacement passed	research-only
Price-space mismatch	direct-price and price-	strong variants were	local policy promoted

Hypothesis	Test	Finding	Disposition
drives the tail	aware refinements	unstable; weak guarded local refinement transferred narrowly	

9.3 Why canonical performance was plausible

The canonical result was initially suspicious because it outperformed several more elaborate methods. The audit did not assume that complexity should win. Four mechanisms make the canonical approach defensible.

First, parity estimates a local forward from a no-arbitrage cross-option relation. Second, expiry-level SVI avoids forcing one parameter path through structurally different maturities. Third, the Jacobian weight maps the variance residual toward price significance without placing Black pricing inside every optimization step. Fourth, repaired-grid projection separates static consistency from quote fit.

Those advantages can coexist with bias. The same chain contributes to carry, fitting, and scoring; a frozen panel influenced policy development; and a repaired control was initially compared with unrepaired candidates. The later audits reduce these specific concerns but do not convert the archive into an independently randomized or long-history study. The conclusion is limited: no tested endogeneity mechanism explains the main canonical advantage on the available panels, and the remaining claims are conditional on those panels.

9.4 Tail behavior changed the research target

Mean RMSE encouraged formulations that fit many ordinary contexts and failed badly on a few. The principal adverse examples were in NDX nonstandard weekly contracts, where raw-price errors can be large. This changed the optimization question from `Which model has the lowest average error?` to `Which intervention improves the identified contexts without creating a new adverse tail?`

The revised diagnostics included lower-tail paired deltas, maximum adverse delta, support overlap, boundary error share, parameter slack, strike-side influence, repair convergence, full-chain repricing, and temporal transfer. A candidate with a lower average but an uncontrolled tail was not promoted.

9.5 Invalid and superseded evidence

Three statuses are kept separate:

- `invalid` means the artifact cannot support a numerical conclusion, as in the contaminated first no-repair run;
- `valid_but_superseded` means the result remains historically interpretable but no longer defines current policy;
- `valid_research_only` means the method was tested correctly but did not meet the evidence required for canonical promotion.

This vocabulary prevents negative outcomes from being rewritten as implementation failures and prevents old controls from silently re-entering current claims.

10. Boundary Conditions, Common Support, and Parameter Fragility

10.1 Boundary hypothesis

Surface tails often fail near the first or last maturity, at sparse strike edges, or outside common support. That was a credible explanation for the adverse weekly contexts and was tested directly.

The principal NDX weekly slice contains 159 quotes. Common-support width is 89.88% of the current log-moneyness width. The slice is neither the first nor last maturity. Boundary quotes represent 2.52% of rows, 0.0128% of fit weight, and 0.1763% of the fit objective. The observed minimum g is positive at 0.2189, and the fitted SVI scale is not at its lower guard. These values do not support endpoint dominance on that target.

The broader attribution reaches the same limited conclusion. In the August NDX focus, boundary squared-error shares range from roughly 0.01% to 0.53% across basket and surface combinations. In the November cross-market panel they remain below 1.46%. Interior common-support observations account for nearly all squared error.

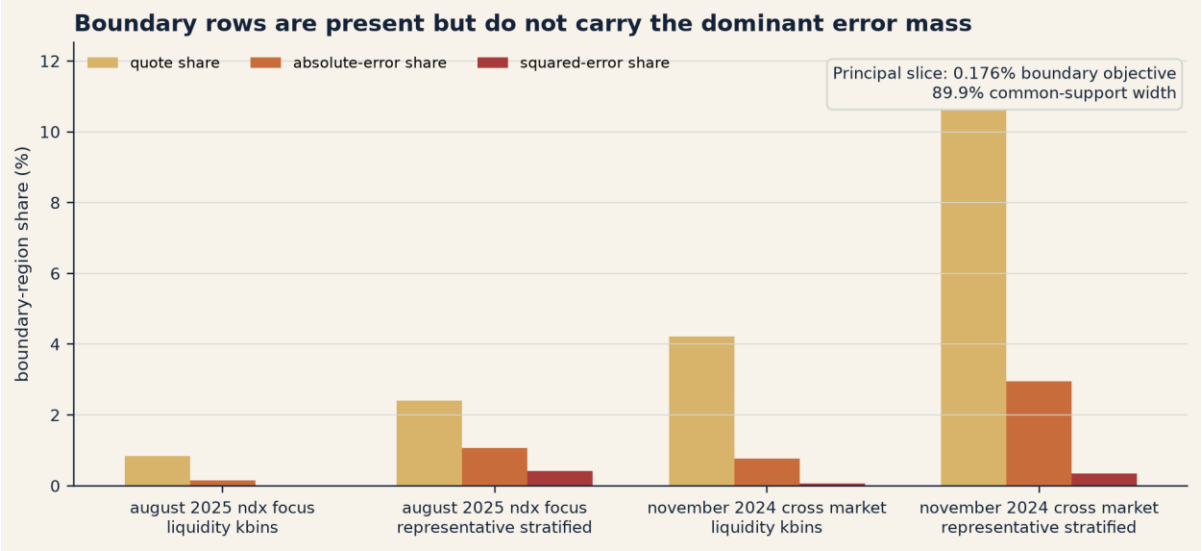

Figure 8. Boundary and common-support attribution. Boundary behavior is secondary in the audited panels; the result is not a claim that boundaries are harmless in all regimes.

10.2 Common support as a treatment definition

If two methods produce prices for different contracts, their RMSE difference mixes model treatment with sample composition. Common support is therefore defined before aggregation. The smaller branch is the relevant denominator when reporting retained support; otherwise, a method that prices very few contracts could appear to have high overlap relative to its own output.

Support matching can also create survivorship bias if difficult contracts fail in both branches. For that reason, common-support RMSE is accompanied by branch quote counts and support ratios. The repo

comparison in Section 13 retains on average 97.48% of the smaller basket, making support loss too small to explain its adverse aggregate result.

10.3 Parameter fragility

The boundary result redirected attention toward cross-sectional influence and parameter fragility. Raw SVI can fit a dense central region while a small strike-side subset moves ρ, m, or σ_svi enough to alter prices across the supported grid. A bounded optimizer does not prevent this; it only prevents parameters from leaving the declared box.

Fragility diagnostics therefore include distance to parameter bounds, minimum g, support width, before/after variance RMSE, price objective, and neighboring calendar compatibility. Strike-side influence caps were tested because one focused NDX date improved when the maximum share was constrained. The transfer tests were mixed and adverse elsewhere, so the cap was not promoted. Its value was diagnostic: it located an influence mechanism without providing a stable global remedy.

10.4 Why Dupire local volatility was not the next repair

Local volatility is a coherent model for matching an arbitrage-free European option surface and generating state-dependent diffusion dynamics. It was not adopted as a solution to the observed tail. Dupire's construction differentiates option prices in strike and maturity. That makes it sensitive to precisely the sparse, noisy, and repair-dependent derivatives under investigation.

More importantly, the diagnosed error was cross-sectional repricing and parameter influence inside a small set of weekly slices. It was not evidence that a deterministic local-volatility dynamics was missing from the pricing of European vanillas. Adding a differentiation-sensitive dynamics layer would increase numerical complexity without identifying the source of the existing residual.

Local volatility remains appropriate future work if the objective changes to exotic pricing or dynamic hedging and if the input surface is sufficiently smooth and arbitrage controlled. It is not a substitute for fixing support, weighting, family routing, or influence in a vanilla calibration benchmark.

11. Failed Regularization and Direct-Price Formulations

11.1 Direct-price objective

The direct-price branch evaluates Black prices inside a bounded least-squares solve. Let P_i(θ) be the model price under SVI parameters θ, M_i the quote midpoint, h_i the half-spread, and θ_0 the canonical starting point. Its residual vector combines

$$r_i(\theta) = \frac{P_i(\theta) - M_i}{\max(h_i, h_{\min})}, \qquad r_{\text{prox}} = \sqrt{\lambda_p}(\theta - \theta_0), \qquad (15)$$

with positive-variance and butterfly penalties on a dense grid. Huber loss limits the influence of large standardized residuals. This is a reasonable research objective: it directly targets price fit while retaining SVI structure and proximity.

The branch often reduced quote-price residuals. It did not pass global promotion because selected configurations failed shape, feasibility, or adverse-tail gates. That is not a contradiction. A flexible

nonlinear fit can lower the objective it is given and still produce a less reliable surface after repair and interpolation.

11.2 Strong price-aware regularization

The project also tested a second variance-stage objective with calibration proximity, price proxies, and parameter-fragility penalties. Strong coefficients sometimes created numerically or economically unstable slices. The optimizer had enough leverage to improve the target region by moving the surface in ways that failed diagnostics elsewhere.

Weak coefficients were safer but often too small to resolve the principal target. Hard guards correctly rejected unsafe changes, yet a guard that always rejects a locally useful adjustment can also be misaligned with the intended objective. This led to explicit guard attribution rather than simply loosening thresholds until a candidate passed.

11.3 Support and residual reweighting

Upweighting contracts inside common support was tested to align fit and evaluation domains. A strength of 0.25 did not resolve the targeted tail, consistent with the later finding that support boundaries carried little objective mass.

Local residual reweighting accepted a small number of cells but left the principal target materially unchanged. It reacted to where the current model was wrong without identifying whether the error arose from price geometry, parameter leverage, or noise. Reweighting the same residuals was insufficient.

11.4 Strike-side cap

A 60% maximum strike-side influence cap improved a focused NDX target. Frozen and November transfer tests were mixed and included adverse behavior. The cap was rejected as a global policy. This is a useful negative result because it separates a diagnostic from a treatment: concentrated side influence existed, but a hard global cap imposed the wrong bias in other contexts.

11.5 What these failures established

The rejected methods narrowed the admissible solution space.

- The main target was not primarily a support-boundary problem.
- A lower direct price objective did not guarantee tail safety after repair.
- Strong global regularization could transfer local error rather than remove it.
- A hard influence cap was too coarse across regimes.
- Weak local adjustment remained plausible if its acceptance rule was explicit and its effect was limited to the diagnosed family.

This sequence motivated the guarded weekly refinement. It is weaker than the failed global formulations and stricter about where it can apply.

12. Guard Attribution and Dual-Objective Local Refinement

12.1 Local refinement

The promoted candidate begins from the canonical raw-SVI slice and performs a weak price-local least-squares adjustment with price and parameter-proximity strengths of 0.25. The adjustment remains inside the SVI parameterization. It does not replace the surface with an unconstrained price interpolant.

The first focused run materially improved the target price objective but failed an existing weighted-variance-RMSE guard. Rather than relaxing the guard, the project attributed the rejection. The guard was protecting the original variance objective, while the candidate was designed to improve spread-scaled price residuals. The final rule retains both objectives: a candidate must make the intended price improvement and remain within explicit variance, parameter, and shape limits.

Eligibility is restricted to true `weekly_pm_nonstandard` contexts. Acceptance checks include quote count, support width, price-objective improvement, variance-RMSE change, ρ movement, minimum σ_{svi}, minimum butterfly g, calendar-neighbor compatibility, and finite output. At most one eligible slice is selected per family context. The choice is made by the declared candidate objective, not by observing final benchmark RMSE.

12.2 Frozen August result

The frozen panel contains 24 SPX, NDX, and RUT asset-date contexts over eight August 2025 dates. Seven contexts change. The other 17 are exact control values because the candidate is ineligible, rejected by a guard, or not selected.

Table 5. Frozen August weekly policy result.

Evaluation basket	Contexts	Control mean RMSE	Guarded mean RMSE	Control maximum	Guarded maximum
Liquidity	24	7.4179	3.8995	87.2883	13.0606
Representative	24	8.6654	3.9681	115.7730	10.9314

The unchanged medians, 2.7162 and 2.3868, are informative. The policy does not shift the center of the panel. It removes a small number of severe tail errors. This is the behavior expected from a guarded local treatment and is why mean improvement alone would be an incomplete description.

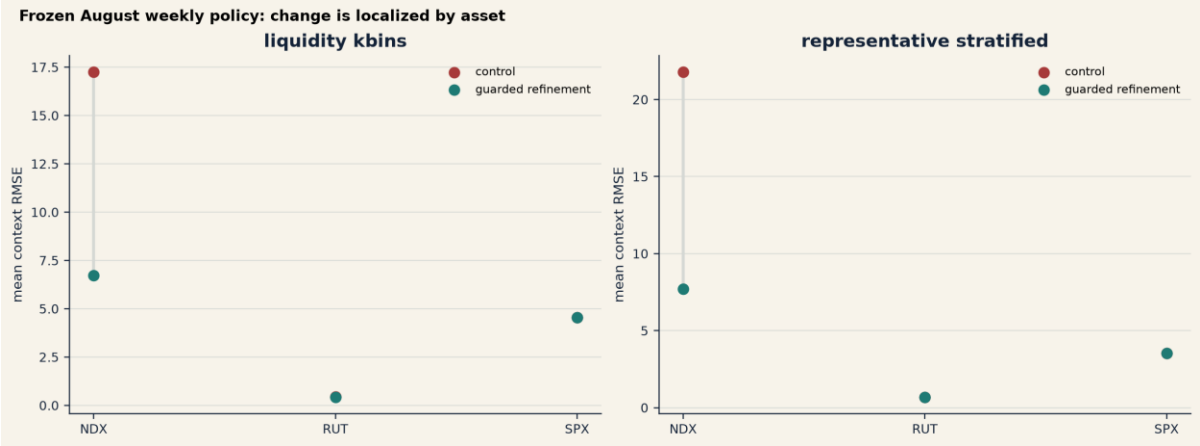

Figure 9. Frozen August weekly result by asset and basket. Raw-price scales differ by asset; the lines are paired within asset-date contexts rather than interpreted as a scale-free pooled model score.

12.3 November transfer

The separate November 2024 panel contains 30 contexts per basket. Mean liquidity RMSE changes from 6.0323 to 5.8766, an improvement of approximately 2.6%. Mean representative RMSE changes from 5.0726 to 4.9719, approximately 2.0%. Medians and maxima are unchanged because the policy activates selectively.

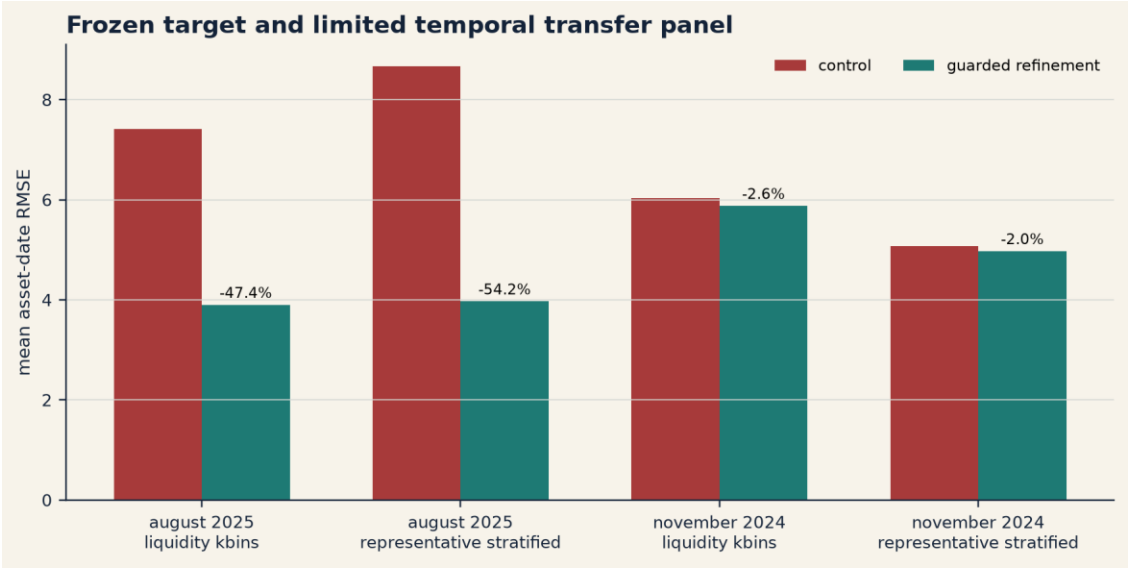

Figure 10. August target and November transfer. The transfer result is directionally consistent but materially smaller. One separate month is not comprehensive temporal validation.

The November result is evidence against an immediate broad regression on that month; it is not evidence of stationarity. The archive does not contain a random sequence of weekly regimes, and the policy was developed after extensive examination of related index panels.

12.4 Full-supported-chain companion

The full-chain study evaluates every contract within supported surface coordinates. After removing duplicated basket labels, it contains 24 independent asset-date contexts. Mean RMSE falls from 6.3036 to 3.1700. None of the seven changed contexts is worse on the full chain.

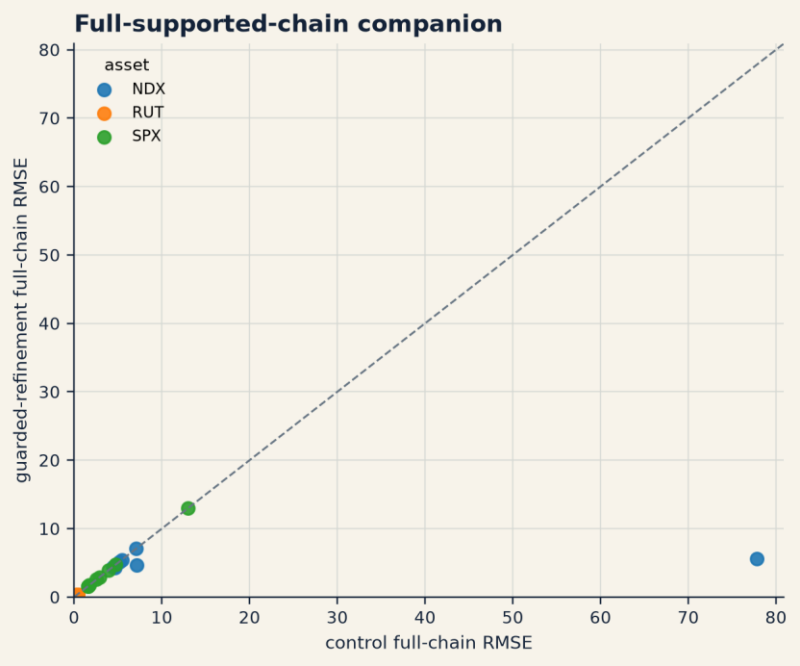

Figure 11. Full-supported-chain companion. Each asset-date chain is counted once; the result is not obtained by duplicating it across export-basket labels.

This check addresses a specific endogeneity concern: the candidate is not merely improving the exported strikes by moving error to unexported contracts inside support. It does not address extrapolated strikes, unobserved dates, or transaction- cost-adjusted economic value.

12.5 Promotion decision

The promoted policy is deliberately narrower than the observed local success. It is not applied to standard monthly, month-end, quarter-end, or American families. It is not a general direct-price refit. It cannot choose among the entire research library on each date. A proposed adjustment that fails any no-harm condition leaves the canonical slice unchanged.

This is an example of governance changing the form of the model. The research did not find a globally better parameterization. It found a localized failure mechanism and a constrained intervention that improved the frozen target, transferred weakly to a separate month, and did not lose on the targeted full chain. That evidence supports a narrow policy and no more.

13. Alternative Surfaces and Repo-Curve Research

13.1 Repair-matched eSSVI

The SSVI family imposes a structured relationship between smile shape and maturity. The implemented eSSVI slice uses

$$w(k) = \frac{\theta}{2}\left[1 + \rho\psi k + \sqrt{(\psi k + \rho)^2 + 1 - \rho^2}\right], \qquad (16)$$

where θ is at-the-money total variance, ρ controls skew, and $\psi = \theta\varphi$ controls curvature. The SLSQP fit enforces the implemented sufficient butterfly conditions $\psi(1+|\rho|) \leq 4$ and $\psi^2(1+|\rho|) \leq 4\theta$.

The first eSSVI comparison was incomplete because repair treatment differed. The final study applies the same post-fit repair wrapper to candidate and canonical grids and evaluates five atomic families in both baskets. Only `monthly_pm_standard` passes paired-mean, tail, diagnostic, and cross-basket gates in both evaluations. Its mean candidate RMSE is 2.2325 against 4.2582 on the liquidity basket and 2.4752 against 4.4376 on the representative basket, with a 92.3% context win rate in both.

Other families fail for different reasons. `weekly_pm_nonstandard` has a favorable mean but an adverse liquidity tail as large as 10.588 in the paired sign convention. `quarter_end_pm` improves average RMSE but fails tail or cross-basket requirements. `eom_pm` and `monthly_am_standard` are adverse on average.

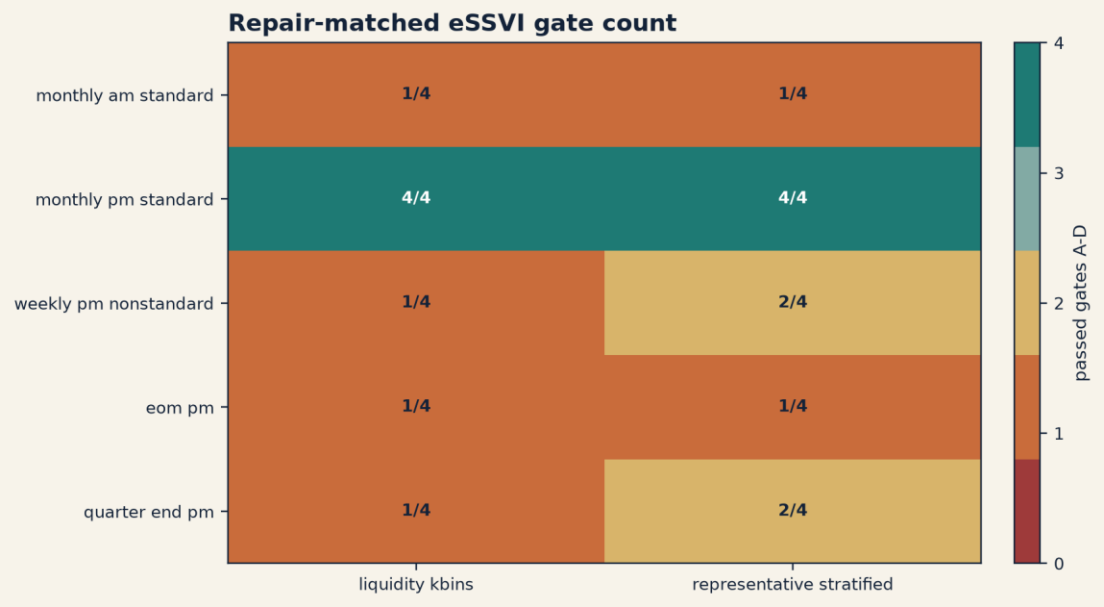

Figure 12. Repair-matched eSSVI gate matrix. The monthly-PM result remains a local candidate; the figure does not support a global eSSVI replacement.

The correct disposition is research candidate. The monthly-PM sample contains 13 paired contexts per basket, and the project has not established temporal breadth or parameter stability sufficient to replace the canonical policy. The result is strong enough to preserve for future testing and too narrow to promote globally.

13.2 Joint SVI

Joint cross-expiry SVI was tested to improve term-structure coherence before repair. It is a natural response to slice independence, but it couples estimation errors and can make one maturity pay for another. Smoke and frozen research panels produced a useful comparator without a stable broad advantage after matched repair and diagnostics. The method remains implemented research code; the report does not imply an exhaustive hyperparameter search or a proof that all joint parameterizations are inferior.

13.3 European repo from parity

The professor-style research branch reframed carry as an implied repo curve. For a given expiry and strike, European parity gives

$$F = K + \frac{C - P}{D}, \qquad q = r - \frac{\log(F/S)}{T}. \qquad (17)$$

This is algebraically consistent with the canonical parity coordinates. The research difference lies in how strikes and expiries are selected, bootstrapped, smoothed, and fed back into surface calibration.

13.4 American volatility-matching repo

For American contracts, European put-call parity is not directly available because exercise rights differ. The research formulation selects a near-forward strike and, for trial repo q, solves a call and put American implied volatility with a CRR tree. The outer root is

$$G(q) = \sigma_C^A(q) - \sigma_P^A(q) = 0. \qquad (18)$$

Each evaluation of G contains two inner bracketed roots. The outer algorithm first scans a declared repo interval to find a sign-changing segment and then applies Brent's method. Failed inner roots do not become outer observations. The implementation therefore rejects the stronger statement that the gap is always monotone and always solvable. Monotonicity may be observed in regular regions; the code requires a numerical bracket rather than assuming it.

The formulation is quantitatively coherent and expensive. It also amplifies quote noise: a small call or put price perturbation changes an inner IV root, and the difference of two roots drives the outer solve. That sensitivity motivated smoothing.

13.5 Local-linear smoothing

Let y_i=log F(T_i). At target maturity T_0, the smoothed curve is the local intercept

$$(\hat{a}, \hat{b}) = \arg\min_{a,b} \sum_i \omega_i(T_0)[y_i - a - b(T_i - T_0)]^2, \qquad \omega_i(T_0) = u_i \exp\left[-\frac{(T_i - T_0)^2}{2h^2}\right]. \qquad (19)$$

Local-linear rather than local-constant smoothing reduces first-order boundary bias and accommodates irregular tenor spacing. Bandwidth remains a modeling choice and sparse support remains a limitation.

13.6 Exact-support repo benchmark

The smoothed repo benchmark name was not accepted as proof of treatment. Curation checked embedded carry metadata for every branch and privately joined parity and local-window prices on exact contract keys. The public output contains only context counts, support ratios, and error sufficient statistics.

Across 202 contexts, mean support relative to the smaller branch is 97.48%, and the local repo wins 80 contexts. Aggregate results remain adverse. On the liquidity basket, parity RMSE is 9.7099 and local-repo RMSE is 16.0983. On the representative basket, the values are 12.2672 and 18.4895.

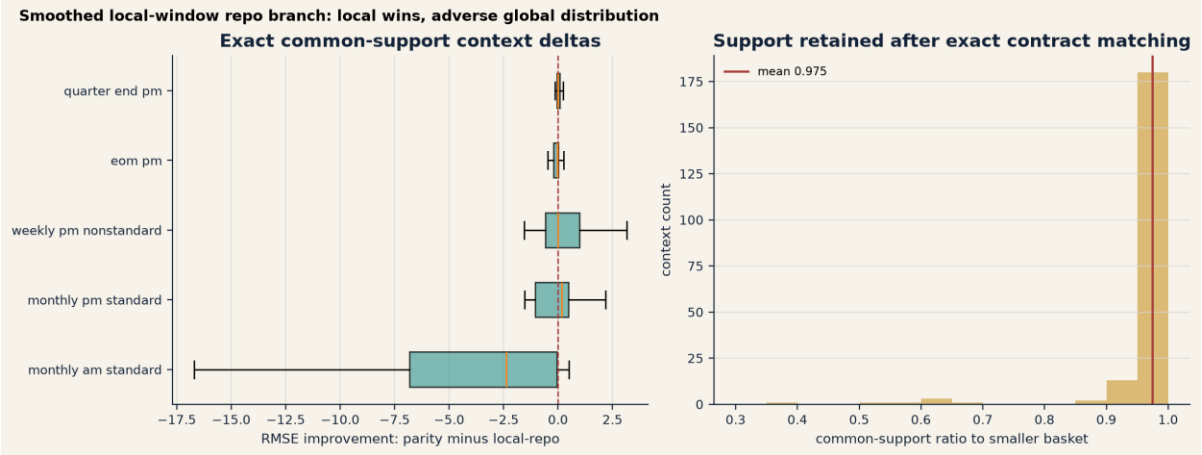

Figure 13. Exact-support local-window repo comparison across all 202 contexts. Positive context deltas favor local repo; aggregate weighted error remains adverse.

The local wins are not discarded. Monthly-PM and selected quarter-end contexts are competitive, and smoothing materially improved coherence over the raw branch. The global result nevertheless rejects a canonical carry replacement. A future study could predeclare family-specific repo eligibility, but it would need genuinely new history and bandwidth sensitivity rather than selecting the favorable contexts from this panel.

14. American Normalization and Discrete-Dividend Treatment

14.1 Separation of normalization and pricing

An American market quote contains both diffusion-related time value and an early- exercise component under the chosen model. Inverting it with a European Black formula assigns both to implied volatility. That is a known approximation in the canonical calibration path. It does not imply that the subsequent price is European: the pricing book uses American CRR and an independent American CN-PSOR check.

The project tested two alternatives. `crr_american_iv` solves the American tree price directly for volatility. `premium_adjusted_black` subtracts an estimated nonnegative early-exercise premium and then inverts the remaining European target, provided the adjusted target lies inside European bounds. A triggered hybrid applies the alternative only when premium is a material share of price and, for calls, when the next dividend is sufficiently near.

14.2 Why premium-heavy puts and ex-dividend calls were targeted

The target was not selected because American effects matter only there. It was selected because those contracts provide the strongest identifying variation for a normalization change. For puts, positive rates and deep-in-the-money states can make immediate exercise valuable. For calls, early exercise is generally tied to dividend economics; a no-dividend American call should equal its European counterpart under the standard model. A blanket American inversion across near-zero-premium quotes adds tree noise where the theoretical adjustment is negligible.

This focus was therefore a power and stability choice. It also limits the result: a triggered rule can miss model error outside the chosen premium threshold or dividend window.

14.3 Empirical normalization result

The saved strict panels cover AAPL, SPY, and INTC. Full American normalization changes some best-surface decisions but does not create a new promotion context. In AAPL, both full challengers are slightly adverse on average and materially worsen two of four contexts. In SPY, full normalization creates one local improvement and two worsenings; it also loses two existing promotion contexts. INTC effects are small and inconsistent across its panels.

Triggered hybrids reduce the scope of change. SPY shows small favorable mean changes under the hybrid rules, but the gains and losses are mixed and no new promotion context appears. Across all packaged summaries, the count of new promotion contexts is zero.

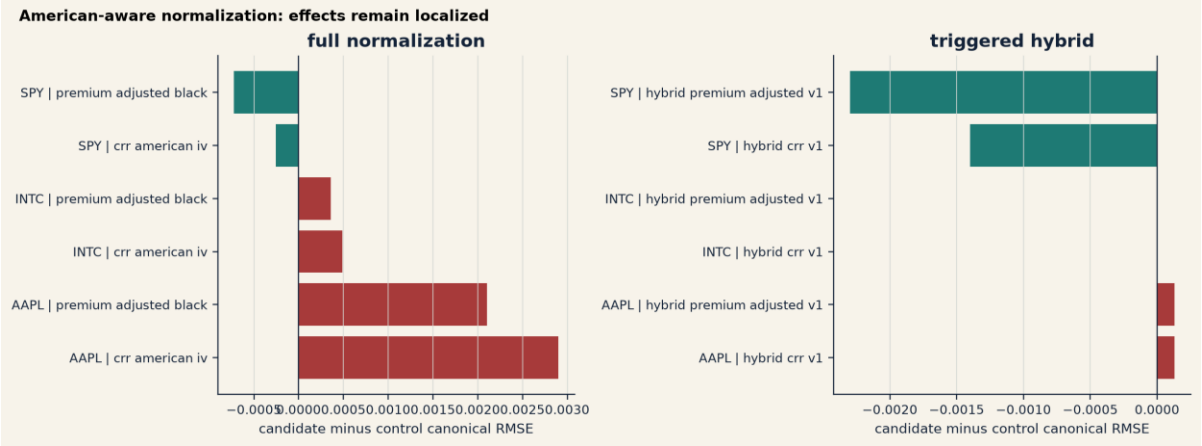

Figure 14. American normalization results. Negative RMSE change favors the challenger. No method establishes a broad replacement of Black normalization.

The baseline is retained because a global change would add model-dependent lattice noise to every American quote without demonstrated downstream benefit. This is a policy decision under limited evidence, not an argument that Black normalization is theoretically exact.

14.4 Discrete-dividend limitation

The auxiliary forward subtracts forecast cash-dividend present value, while the current CRR and PDE APIs receive an equivalent continuous yield. That is a useful interface for broad pricing but does not

place known cash jumps at ex-dividend nodes. The limitation is most relevant for calls whose exercise decision changes immediately before a dividend.

A stronger American engine would insert cash jumps into the tree or PDE grid, reinterpolate the post-dividend value, and allow exercise at the correct event times. It would then require a convergence study around dividend dates and a normalization rerun. That is appropriate future numerical work. It was not added to the frozen policy because the available American panels did not support replacing the simpler normalization, and no new data source is available to expand the study.

14.5 Appropriate output use

American implied volatility in the package should be read as a surface coordinate. American theoretical value, independent-engine disagreement, intrinsic bounds, and early-exercise premium are separate pricing-book fields. A consumer can therefore observe that a quote was normalized by Black but priced by CRR, rather than receiving one number whose convention is implicit.

15. Early-Exercise-Premium Decomposition

15.1 Definition

For identical spot, strike, maturity, rate, carry, volatility, and numerical tree settings, the project defines

$$\mathrm{EEP} = V_A - V_E, \qquad V_A = V_E + \mathrm{EEP}. \qquad (20)$$

This is a matched numerical decomposition. Classical work represents American value as European value plus an integral premium related to the optimal exercise boundary. The implementation does not estimate that integral or claim a structural attribution of each premium dollar. It computes the value of the exercise right under matched CRR inputs.

Matching is essential. If the American and European values use different volatility, carry, grid, or interpolation conventions, their difference mixes exercise with model changes. The public API accepts one input dictionary and changes only the exercise style.

15.2 Numerical conditions

Under a valid arbitrage-free American solver, $\mathrm{V\_A} \geq \mathrm{V\_E}$ because the American holder can choose not to exercise early. A materially negative premium is therefore a diagnostic failure. Tiny negative differences inside numerical tolerance can arise from convergence error and are not interpreted as economic evidence.

The decomposition also reports an identity residual and both engine statuses. A premium is invalid if either matched price fails. For a non-dividend-paying call, the C++ invariant suite verifies equality between European and American CRR values to numerical tolerance.

15.3 Closing-panel distribution

The closing American subset contains 24 asset-date contexts and 1,200 matched pairs from INTC, NVDA, and SPY. Pair coverage is complete. Mean premium is 0.004264, and no negative-premium or intrinsic-bound failure is present.

The average conceals a strongly right-skewed distribution. INTC put premium averages 0.00141 with a maximum of 0.00991. NVDA put premium averages 0.00539, has a 99th percentile of 0.0461, and reaches 0.4172 in one contract. SPY put premium averages 0.01197, with a 99th percentile of 0.1325 and maximum 0.2897. Call premiums are zero or near zero for most observations. Maturities between one and three months have mean premium 0.00751 against 0.00255 below one month in this sample.

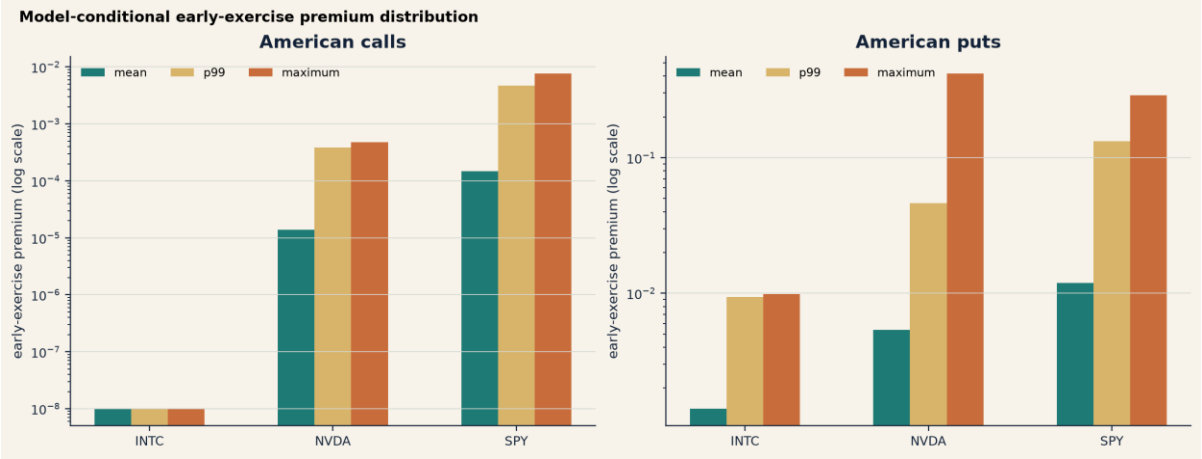

Figure 15. Matched numerical early-exercise premium. The decomposition identifies exercise-option value under the engine; it is not a unique economic attribution.

15.4 Pricing and signal use

EEP is useful as a diagnostic and a feature. It identifies contracts for which European normalization is most likely to be distorted, tests American-engine monotonicity, and separates a surface-price explanation from an exercise explanation. It can also support signals such as premium share of market time value.

It should not be used as a trading signal without additional work. A positive model premium is not mispricing. Execution costs, borrow, exercise operational risk, dividend uncertainty, and model uncertainty can dominate the amount. In the closing panel the mean premium is small relative to typical index pricing errors, while the tail is contract specific.

15.5 Extension path

A future decomposition could include analytic or integral early-exercise-premium representations as independent checks, explicit exercise-boundary output, and cash- dividend event attribution. Those additions would be useful only if the input model and dividend schedule are held fixed across methods. The existing numerical identity is intentionally simpler and more robust to formula-specific assumptions.

16. CRR, Leisen-Reimer, and PDE Numerical Validation

16.1 CRR tree

For time step $\Delta t = T/N$, the Cox-Ross-Rubinstein construction uses

$$u = e^{\sigma\sqrt{\Delta t}}, \qquad d = u^{-1}, \qquad p = \frac{e^{(r-q)\Delta t} - d}{u - d}, \qquad C = e^{-r\Delta t}[pV_u + (1-p)V_d]. \qquad (21)$$

At an American node, the value is $\max(C, \Phi(S))$, where Φ is the intrinsic payoff. The implementation rejects nonfinite inputs and any probability outside $[0,1]$; it does not clip the probability and continue under a modified tree. The Python CRR engine supports controlled batch pricing. The C++ engine provides an independent implementation and test target.

16.2 Leisen-Reimer tree

CRR convergence can oscillate as the terminal strike moves relative to lattice nodes. Leisen-Reimer instead maps Black normal probabilities through the Peizer-Pratt inversion. In the implemented Method 2 notation,

$$h_N(z) = \tfrac{1}{2} + \operatorname{sgn}(z)\sqrt{\tfrac{1}{4} - \tfrac{1}{4}\exp\left[-\left(\frac{z}{N + 1/3 + 0.1/(N+1)}\right)^2 (N + \tfrac{1}{6})\right]}. \qquad (22)$$

The engine promotes an even requested step count to the next odd count, then derives the up probability and stock probability from $h\_N(d\_2)$ and $h\_N(d\_1)$. Probability and denominator guards remain explicit. The C++ tests verify odd-step enforcement and tighter European convergence than the companion CRR configuration in the synthetic case. Leisen-Reimer is maintained as a high-quality alternative, not used as a daily best-tree selector.

16.3 Finite-difference operator

Under continuous carry, the Black-Scholes spatial operator is

$$\mathcal{L}V = \tfrac{1}{2}\sigma^2 S^2 V_{SS} + (r-q)SV_S - rV. \qquad (23)$$

On a uniform spot grid, central differences produce a tridiagonal system. The theta step is

$$[I - \theta\Delta t\,\mathcal{L}_h]V^n = [I + (1-\theta)\Delta t\,\mathcal{L}_h]V^{n+1} + b^n, \qquad (24)$$

where $\theta = 1$ gives the fully implicit scheme and $\theta = 1/2$ gives Crank-Nicolson. Boundary values depend on option type, exercise style, remaining maturity, rate, and carry. The European tridiagonal systems are solved directly.

16.4 American complementarity and PSOR

American exercise produces a discrete linear-complementarity problem. In continuous notation,

$$V - \Phi \geq 0, \qquad -(V_t + \mathcal{L}V) \geq 0, \qquad (V - \Phi)\,[-(V_t + \mathcal{L}V)] = 0. \qquad (25)$$

The C++ CN-PSOR solver applies projected successive over-relaxation to the tridiagonal step and projects each iterate onto intrinsic value. Relaxation must lie strictly between 0 and 2. Failure to meet the iteration tolerance is returned as a solver failure rather than a stale price.

The terminal payoff is nonsmooth. Pure Crank-Nicolson can propagate initial oscillations on a fine grid. The implementation uses Rannacher startup: the first configured intervals are split into implicit half steps before returning to $\theta=1/2$. In operator form,

$$V^{n+1/2} = \left[I - \tfrac{\Delta t}{2}\mathcal{L}_h\right]^{-1} V^{n+1}, \qquad V^{n} = \left[I - \tfrac{\Delta t}{2}\mathcal{L}_h\right]^{-1} V^{n+1/2}. \qquad (26)$$

The American projection is applied at each half step.

16.5 Convergence study

The packaged convergence panel contains two anonymized SPY call contracts from one date. CRR uses 3,200 steps as its family reference. The 400-step production configuration differs by at most 0.00138 from that reference. The sequence is not monotone: the 200- and 400-step errors can exceed the 100-step error. This is expected lattice behavior and one reason a single coarse-versus-fine comparison is insufficient.

The original 400-by-400 CN-PSOR grid at spot scale 5 is materially under-resolved: absolute differences to the CRR reference are approximately 0.323 and 0.0488. Grid refinement reduces the error, but spot-domain scale also matters. The promoted independent configuration is 1,600 time steps by 1,600 spot steps with maximum spot equal to four times current spot, two Rannacher startup intervals, relaxation 1.2, and tolerance 10^{-8}. Its differences to the CRR reference are approximately 0.00524 and 0.000605 on the two contracts.

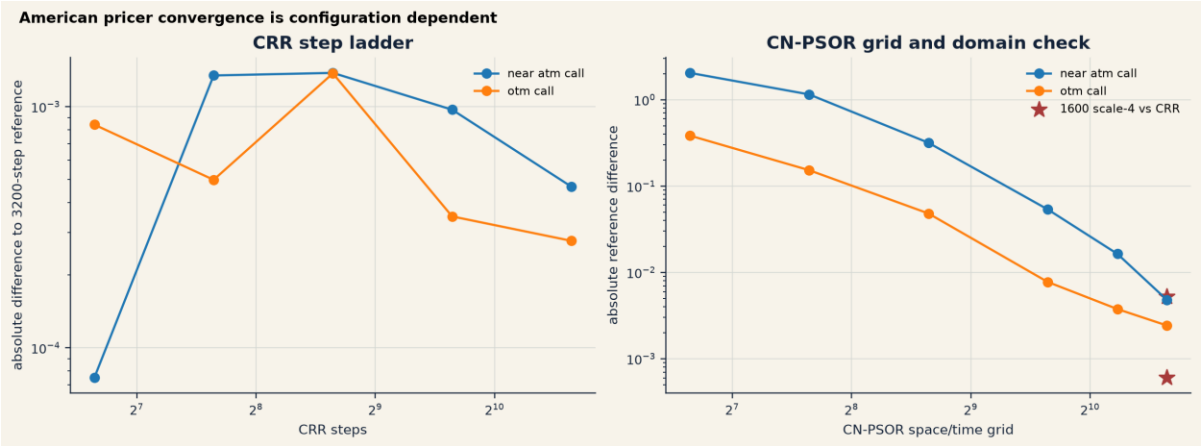

Figure 16. CRR and CN-PSOR convergence ladders. Family reference rows are retained in the table but omitted as artificial zero-error endpoints in the plotted ladders. Neither family reference is ground truth.

The panel is too small to establish a universal PDE grid. It is sufficient to reject the earlier coarse setting and to document a materially more resolved independent check. A broader convergence design would stratify option side, maturity, moneyness, dividend timing, volatility, and spot-domain truncation.

16.6 C++ library state

The C++17 library exports namespaced inputs and result types, validates all economic and numerical inputs, and installs CMake package configuration for external use. Six CTest suites cover Black pricing, trees, European finite differences, American pricing, invariants, and invalid inputs. The public restructure

removed the private market-file dependency from the executable and verified an external `find_package` consumer.

The library is a maintained numerical component. It is not yet called as the policy PDE engine from the Python snapshot pipeline. That separation is explicit in the release manifest and should not be described as an integrated production cross-check until the invocation and result transport are implemented.

17. Closing Cross-Market Results

17.1 Panel and aggregation

The closing bundle uses six assets over eight August 2025 dates: INTC, NDX, NVDA, RUT, SPX, and SPY. It contains 48 asset-date contexts and 2,400 valid canonical prices. Each asset contributes 400 prices. European indices use Black pricing from the canonical repaired surface. American equity and ETF contracts use CRR under American exercise.

For errors e_i=P_i-M_i, the bundle summaries are

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}, \qquad \mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |e_i|. \qquad (27)$$

The pooled RMSE is 3.5122, MAE is 1.4620, and maximum absolute error is 27.7368. Those values are recombined from context sufficient statistics with row-count weights. They are not scale-free performance measures.

Table 6. Closing pricing error by asset.

Asset	Style	Engine	Contexts	Prices	RMSE	MAE	Maximum absolute error
INTC	American	CRR	8	400	0.0540	0.0228	0.6321
NDX	European	Black	8	400	7.9061	5.7608	27.7368
NVDA	American	CRR	8	400	0.1525	0.0836	1.3260
RUT	European	Black	8	400	0.6362	0.4426	3.7832
SPX	European	Black	8	400	3.3184	2.2724	14.7489
SPY	American	CRR	8	400	0.2543	0.1899	1.2180

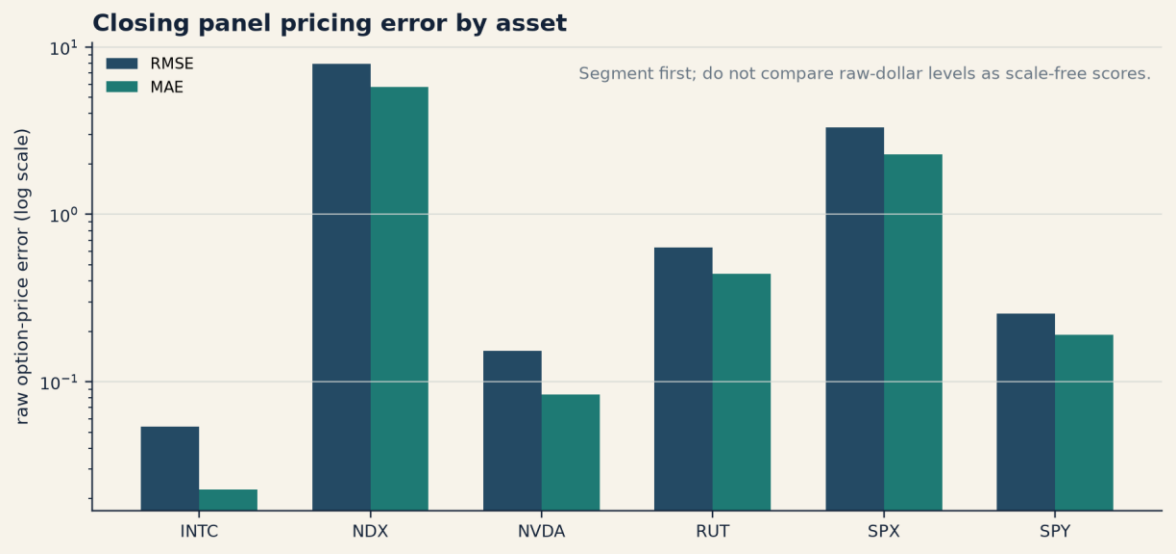

Figure 17. Per-asset closing errors on a log scale. The display prevents the index price scale from visually suppressing equity and ETF errors.

NDX dominates the raw-dollar aggregate. The table does not establish that NDX is the least accurate asset in relative or spread-normalized terms because those metrics answer a different question. It establishes where the closing pricing-book dollar error resides.

17.2 Paired and common-support statistics

For surface or carry treatment m, the context-level paired improvement is

$$\Delta_j = \mathrm{RMSE}_{\mathrm{control},j} - \mathrm{RMSE}_{m,j}, \qquad (28)$$

so positive values favor the candidate. For branches with possible price failures, the support ratio is

$$R_j = \frac{|Q_{0,j} \cap Q_{m,j}|}{\min(|Q_{0,j}|, |Q_{m,j}|)}. \qquad (29)$$

These definitions are repeated here because they govern interpretation of the weekly, eSSVI, and repo results. The closing panel itself is a canonical accounting run rather than a candidate selection exercise.

17.3 Quote and combined quality

Quote-quality pass, warning, and failure rates are 73.92%, 25.79%, and 0.2917%. Seven quotes fail. All failures have relative spread exactly 0.40, the strict limit. Six are classified `liquidity_fail`; one also carries `liquidity_warn`. There are no engine-disagreement failures and no below-intrinsic rows.

Combined quality is stricter than quote quality. Only 13.29% of rows pass combined status, 86.42% warn, and 0.2917% fail. The warning rate is dominated by `surface_diag_warn`, which applies to 2,000 rows, and by 388 engine-disagreement warnings. A warning is not a failed price. It means at least one configured diagnostic requires review.

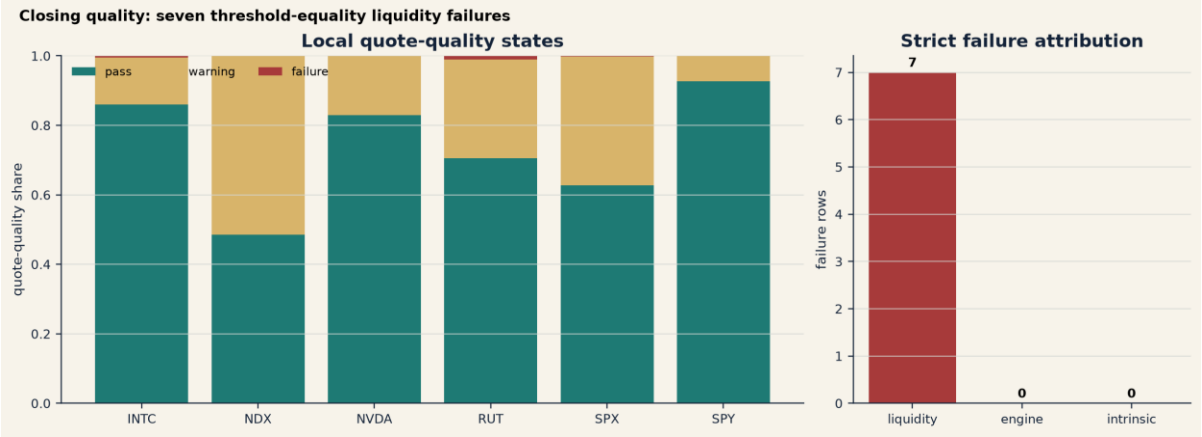

Figure 18. Closing quality and failure attribution. The strict zero-failure release gate remains failed; no threshold is relaxed for presentation.

The deployment validator therefore returns one failure and two warnings. The failure is a data-quality gate, not an engine breakdown. This distinction is operationally important: all 2,400 prices are valid, but the bundle is not eligible for an unqualified zero-failure release.

17.4 Signals and alerts

The closing layer emits 414 signal rows and 61 alerts: 38 `review`, 23 `watch`, and zero `high`. Signal families include 30-day ATM volatility, 25-delta risk reversal and fly, term slope, pricing RMSE, and mean engine disagreement. Counts are an operational inventory, not evidence of trading profitability.

Signals inherit the same surface, carry, and quality state as the pricing book. A deployment consumer can filter on quality status rather than recomputing whether a surface or engine warning should invalidate a feature. The current project stops at feature and alert generation; it does not backtest positions, turnover, slippage, hedging, or capital use.

17.5 What the closing run demonstrates

The closing run demonstrates that the frozen policy can execute across European index and American equity/ETF contexts, produce valid style-correct prices, compute matched premiums, run independent checks, and retain strict quality failures. It does not prove that the theoretical values are unbiased, executable, or stable outside the available dates. The closing panel is a systems and evidence result, not a market- efficiency test.

18. Limitations, Frozen Decisions, and Appropriate Future Work

18.1 Data limitations

The archive is irregular and concentrated in a small number of windows. It is not a balanced panel, a random regime sample, or a long historical backtest. Several key results rely on August 2025 index dates, one November 2024 transfer month, or two SPY convergence contracts. Bootstrap intervals quantify variation across the available paired contexts; they do not correct the sampling design.

Raw vendor data cannot be redistributed, and access is no longer available for refresh. The public release can reproduce every published aggregate and synthetic calculation but cannot independently reconstruct the source chains. Any future empirical expansion should begin with a new licensed or redistributable dataset and a predeclared holdout design rather than extending the existing selected windows.

18.2 Model limitations

Raw SVI is a parsimonious static representation. Repaired-grid projection enforces configured discrete constraints only on the evaluated support. Extrapolation is not validated as globally arbitrage free. The Jacobian objective is a first-order price alignment, not a full likelihood. Quote dependence, microstructure dynamics, and uncertainty in spot, rate, and dividend inputs are not jointly estimated.

American quote normalization remains European Black. Downstream pricing is American, but the surface coordinate can absorb exercise premium. Equivalent continuous yield does not reproduce cash-dividend jumps at exact event times. The numerical EEP is a matched model difference, not a unique structural decomposition.

18.3 Numerical limitations

CRR, Leisen-Reimer, and CN-PSOR are independently implemented, but the policy PDE engine is not yet invoked by the Python snapshot pipeline. The convergence panel is too small to establish one universal grid across all American contracts. PDE domain truncation, mesh concentration, dividend jumps, and exercise-boundary resolution remain contract dependent.

The repaired-grid operator alternates a convex isotonic projection with nonlinear row constraints. Final diagnostics are checked, but no claim is made that the alternation finds a unique global minimum of one joint objective. SLSQP and L-BFGS-B success statuses are treated as numerical evidence only when independent guards pass.

18.4 Experimental limitations

The research library was developed against the available archive. Even with fixed daily policies and common-support audits, model-development endogeneity remains. The November panel is limited, and family-specific sample sizes can be small. The monthly-PM eSSVI candidate and local repo wins are hypotheses for new data, not instructions to mine the current archive more aggressively.

Promotion gates are versioned project controls. They reflect risk preferences about mean improvement, adverse tails, diagnostics, and cross-basket transfer. Another use case can choose different tolerances, but it should disclose and sensitivity-test them rather than inheriting the numbers as universal standards.

18.5 Frozen decisions

The following decisions define version 1.0 of the public research system.

- Five atomic European index families are calibrated separately; `pooled` is reporting only.
- European carry uses expiry-level parity. Smoothed local-window repo remains research only.

- Raw SVI with Jacobian-consistent weighting is the canonical slice fit.
- European atomic families use repaired grids. A failed repair is a failure, not a raw-grid fallback under a repaired label.
- Guarded weak price-local refinement is enabled only for eligible nonstandard weekly contexts.
- American calibration retains Black normalization. American pricing uses CRR with matched European CRR for EEP.
- Leisen-Reimer and CN-PSOR are maintained independent numerical methods, not daily best-engine selectors.
- The strict release gate remains failed when any quote equals the 40% spread limit.

18.6 Appropriate future work

The highest-value future work is evidence expansion, not another in-sample surface variant. A new study should use a predeclared set of assets, dates, families, and holdouts; retain common-support and full-chain companions; and freeze thresholds before scoring. The monthly-PM eSSVI candidate and family-specific repo behavior are appropriate first challengers because they already have a defined hypothesis.

On the American side, the next numerical improvement is explicit cash-dividend event handling in tree and PDE engines, followed by a stratified convergence study. That would make ex-dividend call exercise and premium decomposition more defensible. A Python-to-C++ policy integration should then transport full engine metadata and fail closed on invocation or convergence errors.

Local volatility is appropriate only if the project moves toward exotic pricing or dynamic hedge analysis and can provide a sufficiently smooth arbitrage-controlled surface. It is not the next fix for the existing vanilla weekly tail. Signal research should begin only after defining transaction costs, execution timing, hedge policy, and a genuine holdout; otherwise the signal layer remains descriptive.

18.7 Final state

The project is not a failed pricing model. It is a bounded research result with a functional reference implementation, a narrow promoted surface intervention, an explicit American pricing stack, and a documented set of non-promoted alternatives. Its empirical scope is not sufficient for unattended deployment or broad market claims. Freezing it in that state is preferable to extending a selected archive until another method passes.

References

Ayer, M., Brunk, H. D., Ewing, G. M., Reid, W. T., and Silverman, E. (1955). "An Empirical Distribution Function for Sampling with Incomplete Information." *Annals of Mathematical Statistics* 26(4), 641-647. https://doi.org/10.1214/aoms/1177728423.

Black, F. (1976). "The Pricing of Commodity Contracts." *Journal of Financial Economics* 3(1-2), 167-179. https://doi.org/10.1016/0304-405X(76)90024-6.

Black, F., and Scholes, M. (1973). "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81(3), 637-654. https://doi.org/10.1086/260062.

Breeden, D. T., and Litzenberger, R. H. (1978). "Prices of State-Contingent Claims Implicit in Option Prices." *Journal of Business* 51(4), 621-651. https://doi.org/10.1086/296025.

Brennan, M. J., and Schwartz, E. S. (1978). "Finite Difference Methods and Jump Processes Arising in the Pricing of Contingent Claims: A Synthesis." *Journal of Financial and Quantitative Analysis* 13(3), 461-474. https://doi.org/10.2307/2330152.

Brent, R. P. (1971). "An Algorithm with Guaranteed Convergence for Finding a Zero of a Function." *Computer Journal* 14(4), 422-425. https://doi.org/10.1093/comjnl/14.4.422.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). "A Limited Memory Algorithm for Bound Constrained Optimization." *SIAM Journal on Scientific Computing* 16(5), 1190-1208. https://doi.org/10.1137/0916069.

Carr, P., Jarrow, R., and Myneni, R. (1992). "Alternative Characterizations of American Put Options." *Mathematical Finance* 2(2), 87-106. https://doi.org/10.1111/j.1467-9965.1992.tb00040.x.

Cox, J. C., Ross, S. A., and Rubinstein, M. (1979). "Option Pricing: A Simplified Approach." *Journal of Financial Economics* 7(3), 229-263. https://doi.org/10.1016/0304-405X(79)90015-1.

Crank, J., and Nicolson, P. (1947). "A Practical Method for Numerical Evaluation of Solutions of Partial Differential Equations of the Heat-Conduction Type." *Proceedings of the Cambridge Philosophical Society* 43(1), 50-67. https://doi.org/10.1017/S0305004100023197.

Cryer, C. W. (1971). "The Solution of a Quadratic Programming Problem Using Systematic Overrelaxation." *SIAM Journal on Control* 9(3), 385-392. https://doi.org/10.1137/0309028.

Dupire, B. (1994). "Pricing with a Smile." *Risk* 7(1), 18-20.

Efron, B. (1979). "Bootstrap Methods: Another Look at the Jackknife." *Annals of Statistics* 7(1), 1-26. https://doi.org/10.1214/aos/1176344552.

Fan, J. (1992). "Design-Adaptive Nonparametric Regression." *Journal of the American Statistical Association* 87(420), 998-1004. https://doi.org/10.1080/01621459.1992.10476255.

Fritsch, F. N., and Carlson, R. E. (1980). "Monotone Piecewise Cubic Interpolation." *SIAM Journal on Numerical Analysis* 17(2), 238-246. https://doi.org/10.1137/0717021.

Gatheral, J., and Jacquier, A. (2014). "Arbitrage-Free SVI Volatility Surfaces." *Quantitative Finance* 14(1), 59-71. https://doi.org/10.1080/14697688.2013.819986.

Hendriks, S., and Martini, C. (2019). "The Extended SSVI Volatility Surface." *Journal of Computational Finance* 22(5), 25-39. https://doi.org/10.21314/JCF.2019.365.

Huber, P. J. (1964). "Robust Estimation of a Location Parameter." *Annals of Mathematical Statistics* 35(1), 73-101. https://doi.org/10.1214/aoms/1177703732.

Kim, I. J. (1990). "The Analytic Valuation of American Options." *Review of Financial Studies* 3(4), 547-572. https://doi.org/10.1093/rfs/3.4.547.

Kraft, D. (1988). *A Software Package for Sequential Quadratic Programming*. DFVLR-FB 88-28, Deutsche Forschungs- und Versuchsanstalt fuer Luft- und Raumfahrt.

Leisen, D. P. J., and Reimer, M. (1996). "Binomial Models for Option Valuation: Examining and Improving Convergence." *Applied Mathematical Finance* 3(4), 319-346. https://doi.org/10.1080/13504869600000015.

Merton, R. C. (1973). "Theory of Rational Option Pricing." *Bell Journal of Economics and Management Science* 4(1), 141-183. https://www.jstor.org/stable/3003143.

Rannacher, R. (1984). "Finite Element Solution of Diffusion Problems with Irregular Data." *Numerische Mathematik* 43(2), 309-327. https://doi.org/10.1007/BF01390130.

Stoll, H. R. (1969). "The Relationship Between Put and Call Option Prices." *Journal of Finance* 24(5), 801-824. https://doi.org/10.1111/j.1540-6261.1969.tb01694.x.

Whaley, R. E. (1981). "On the Valuation of American Call Options on Stocks with Known Dividends." *Journal of Financial Economics* 9(2), 207-211. https://doi.org/10.1016/0304-405X(81)90013-1.

Appendix A. Notation and Equation Index

A.1 Notation

Symbol	Definition
S	current underlying spot price
K	option strike
T	year fraction from observation to expiry
$D(T)$	discount factor to expiry
$r(T)$	continuously compounded zero rate, $-\log D(T)/T$
$F(T)$	forward price under the declared carry state
$q(T)$	equivalent continuous repo or dividend yield
C,P	call and put values
ω	option sign, +1 call and -1 put
σ	Black implied volatility or model volatility
k	log-forward moneyness, $\log(K/F)$
w	total implied variance, $\sigma^2 T$
a,b,ρ,m,σ_{svi}	raw-SVI parameters
$g(k)$	SVI density or butterfly diagnostic function
h_i	quote half-spread after configured floor
λ_i	normalized calibration weight
$\Phi(S)$	intrinsic option payoff
V_E,V_A	matched European and American model values
EEP	early-exercise premium, V_A-V_E
$Q_{m,j}$	contract set priced by method m in context j
Δ_j	paired context improvement, positive for candidate

A.2 Equation traceability

Equation 6 is the displayed root equation; the style-correct bounds mapped by EQ-006 are stated directly before it. Equations 9 and 10 are the chain-rule derivation and its squared weight, both covered by EQ-010.

Table A1. Displayed-equation index. Full paths, source identifiers, and audit methods are in `report/equation_traceability.csv`.

Eq.	Audit ID	Topic	Primary implementation symbol
1	EQ-001	parity regression	`fit_european_parity`
2	EQ-002	discount and forward recovery	`fit_european_parity`
3	EQ-003	cash-dividend forward	`auxiliary_cash_dividend_carry`
4	EQ-004	forward coordinates	`svi_total_variance`
5	EQ-005	Black forward price	`black_forward_price`; `black_scholes`
6	EQ-007	implied-volatility root	`solve_monotone_volatility`
7	EQ-008	raw SVI	`svi_total_variance`
8	EQ-009	SVI butterfly function	`butterfly_g_from_values`
9-10	EQ-010	Jacobian weight derivation	`jacobian_consistent_weights`
11	EQ-011	bounded SVI objective	`calibrate_svi_slice`
12	EQ-012	calendar isotonic projection	`isotonic_increasing`
13	EQ-013	discrete butterfly proxy	`discrete_butterfly_g`
14	EQ-014	alternating grid repair	`repair_surface_grid`
15	EQ-016	direct-price residual	`fit_svi_direct_price`
16	EQ-015	eSSVI slice	`essvi_total_variance`
17	EQ-017	European implied repo	`european_implied_repo`
18	EQ-018	American volatility-matching repo	`fit_american_volatility_matching_repo`
19	EQ-019	local-linear carry smoother	`smooth_log_forward_curve`
20	EQ-026	early-exercise premium	`decompose_early_exercise`
21	EQ-020	CRR transition	`price_crr`; `cox_ross_rubinstein`
22	EQ-021	Leisen-Reimer probability map	`peizer_pratt_method2`; `leisen_reimer`
23	EQ-022	Black-Scholes PDE operator	`build_theta_matrix`
24	EQ-023	theta time step	`build_theta_rhs`
25	EQ-024	American complementarity	`projected_sor`; `finite_difference`
26	EQ-025	Rannacher startup	`finite_difference`
27	EQ-027	RMSE and MAE	`summarize_pricing_book`
28	EQ-028	paired candidate delta	`paired_comparison`
29	EQ-029	common-support ratio	`support_attribution`
30	EQ-030	guarded weekly objective	`refine_weekly_slice`

Appendix B. Algorithm Pseudocode

B.1 European parity carry

```
INPUT common-strike European calls and puts for one expiry
REQUIRE at least the configured number of valid call-put pairs
y <- call_mid - put_mid
fit y = alpha + beta * strike
discount <- -beta
forward <- alpha / discount
rate <- -log(discount) / maturity
REJECT nonpositive discount, nonpositive forward, or failed diagnostics
RETURN forward, discount, rate, pair count, residual diagnostics
```

B.2 Implied volatility

```
INPUT market price, style-correct price bounds, model price function
REJECT market price outside bounds beyond tolerance
SET bracket = [sigma_min, sigma_max]
IF model residual has no sign change:
    expand upper endpoint through the configured sequence
IF no sign-changing bracket exists:
    RETURN typed bracket failure
SOLVE residual(sigma) = 0 with Brent's method
RETURN sigma, residual, bracket, iterations, convergence status
```

B.3 SVI calibration

```
INPUT accepted k, total variance, positive normalized weights
GENERATE deterministic parameter starts
FOR each start:
    minimize bounded weighted total-variance error with L-BFGS-B
    retain finite successful result
SELECT lowest successful objective
EVALUATE dense-support variance, parameter, and butterfly diagnostics
REJECT if any required guard fails
RETURN parameters, objective, RMSE, support, and diagnostics
```

B.4 Repaired grid

```
INPUT regular total-variance grid
FOR pass = 1 to max_passes:
    project every strike column onto nondecreasing maturity values with PAVA
    for every violating maturity row:
        minimize squared displacement subject to positivity and discrete g constraints
    compute combined calendar and butterfly diagnostics
    IF all diagnostics pass: RETURN converged repaired grid
RETURN failed repair with final grid, displacement, and diagnostics
```

B.5 Guarded weekly refinement

For eligible nonstandard weekly slices, the local candidate solves

$$\min_{\theta\in\Theta}\sum_i\left[\frac{P_i(\theta)-M_i}{\max(h_i,h_{\min})}\right]^2+\lambda_p\|\theta-\theta_0\|_2^2, \qquad (30)$$

then applies the predeclared acceptance rule.

```
INPUT canonical SVI slice and quote/pricing state
REQUIRE family = weekly_pm_nonstandard
REQUIRE maturity, quote count, and support-width eligibility
SOLVE weak price-local SVI adjustment from canonical parameters
```

```
COMPUTE before/after price objective and variance diagnostics
REJECT unless price objective improves
REJECT if variance RMSE, rho slack, SVI sigma, butterfly g,
          calendar neighbor, support, or finite-output guard fails
RANK accepted eligible slices by declared candidate objective
APPLY at most one slice in the family context
RETURN applied or unchanged canonical state with reason codes
```

B.6 American volatility-matching repo

```
INPUT spot, strike, maturity, rate, call price, put price, repo interval
FOR each repo scan point q:
    solve American call IV by bracketed CRR inversion
    solve American put IV by bracketed CRR inversion
    IF both converge: store gap(q) = call_IV - put_IV
FIND adjacent valid scan points with opposite gap signs
IF none exists: RETURN not bracketed
SOLVE gap(q) = 0 with Brent's method
RETURN q, forward, call IV, put IV, residual, evaluation count
```

B.7 Early-exercise decomposition

```
INPUT one contract state and one tree configuration
american <- CRR(input, American)
european <- CRR(input, European)
IF either price fails: RETURN invalid decomposition
premium <- american - european
identity_residual <- american - (european + premium)
FLAG materially negative premium or intrinsic-bound failure
RETURN both prices, premium, residual, and statuses
```

B.8 CN-PSOR

```
INITIALIZE terminal payoff on a bounded spot grid
APPLY configured implicit Rannacher half steps with exercise projection
FOR remaining time levels:
    construct Crank-Nicolson tridiagonal system and boundary contribution
    iterate projected SOR:
        update each interior node with relaxation
        project node value to max(iterate, intrinsic)
        stop when maximum update is below tolerance
    IF iteration limit is reached: RETURN solver failure
INTERPOLATE value at current spot
RETURN price, effective grid, and iteration diagnostics
```

Appendix C. Promotion Gates and Governance Definitions

C.1 Context and support

A comparison unit is an asset, observation date, atomic contract family, and evaluation basket. Candidate and control must use the same exercise style and pricing engine. Carry is held fixed unless carry itself is the treatment. Quote support must be identical or explicitly joined and reported with Equation 29.

C.2 Gate interpretation

- **Paired mean gate:** candidate mean context delta must meet the declared direction and paired uncertainty rule.
- **Tail gate:** lower-tail and maximum adverse deltas must remain inside declared limits. A favorable mean cannot compensate for an uncontrolled date-family loss.
- **Diagnostic gate:** repair, surface, parameter, pricing, and engine failure rates cannot worsen beyond policy.
- **Cross-basket gate:** the decision must be evaluated on liquidity and representative baskets, with a full-supported-chain companion where applicable.
- **Atomic-family gate:** pooled reporting cannot supply or rescue a family promotion.
- **Strict-selection gate:** the policy cannot choose a daily best surface or engine after benchmark observation.

Gate thresholds belong to governance version `2026-04-13-rigor-v1`. They are not presented as universal statistical constants.

C.3 Frozen numerical settings

Table C1. Selected public policy settings.

Component	Setting
Day-count basis	365.25
Black IV root	$[10^{-6},3]$, expand to 5 and 8, tolerance 10^{-8}, 80 iterations
SVI support grid	81 points on $[-0.5,0.5]$
Minimum quotes per SVI slice	5
SVI multistarts	12
Repair	4 passes, 300 SLSQP iterations, tolerance 10^{-8}
Weekly family	`weekly_pm_nonstandard` only
Weekly strengths	price 0.25, proximity 0.25
Weekly maturity interval	0.015 to 0.12 years

Component	Setting
Weekly minimum quotes/support width	12 / 0.05 log-moneyness
Weekly parameter guards	rho slack 0.005; SVI sigma 0.001; minimum g=-0.02
American canonical tree	CRR, 400 steps
Independent American PDE	1600 by 1600, spot scale 4, PSOR 1.2, tolerance 10^{-8}
PDE startup	2 Rannacher intervals
Relative-spread warning/failure	0.25 / 0.40
Strict deployment	zero failures required

C.4 Threshold equality

The relative-spread failure rule is inclusive at 0.40. Seven closing quotes equal the limit and remain failures. Changing equality semantics after seeing the result would be an unrecorded gate change.

Appendix D. Experiment Registry and Rejected-Model Table

D.1 Full registered sequence

Table D1. Frozen experiment dispositions. Detailed hypotheses, panels, evidence, outcomes, and notes are in `experiment_registry.csv`.

ID	Stage	Decision	Validity	Decision class
EXP-001	legacy control	retain historical control	valid but superseded	retained/control
EXP-002	taxonomy	promote classifier	valid	promoted
EXP-003	objective weighting	promote weighting	valid	promoted
EXP-004	repair fairness	exclude	invalid	rejected
EXP-005	repair fairness	retain corrected control	valid	retained/control
EXP-006	repair implementation	promote fix	valid	promoted
EXP-007	root-cause rerun	promote reference	valid	promoted
EXP-008	weekly support	promote fix	valid	promoted
EXP-009	alternative surface	rerun with matched repair	diagnosis only	continue
EXP-010	alternative surface	retain research candidate	valid	retained
EXP-011	alternative surface	retain research only	limited panel	retained
EXP-012	price objective	reject global promotion	valid	rejected
EXP-013	repo inference	add smoothing and retest	raw-method evidence	continue
EXP-014	repo inference	retain research only	valid	retained
EXP-015	American fairness	retain baseline	valid	retained
EXP-016	American normalization	continue to stage 2	valid	continue
EXP-017	American normalization	retain Black baseline	valid	retained
EXP-018	weekly influence	test transfer	focused evidence	continue
EXP-019	weekly influence	reject global promotion	valid	rejected
EXP-020	boundary diagnosis	reject primary-boundary hypothesis	valid	rejected hypothesis
EXP-021	price-aware variance	reject strong variant	valid	rejected
EXP-022	price-aware variance	do not promote	valid	rejected
EXP-023	support weighting	reject	valid	rejected
EXP-024	residual weighting	reject as solution	valid	rejected
EXP-025	weekly price refinement	diagnose guard	focused evidence	continue
EXP-026	guard attribution	advance to frozen transfer	valid	advanced
EXP-027	frozen candidate	promote narrow policy	valid	promoted
EXP-028	temporal transfer	pass limited transfer check	limited panel	advanced
EXP-029	basket endogeneity	require companion full chain	valid	promoted governance
EXP-030	carry endogeneity	carry not primary cause	limited design	retained
EXP-031	American numerics	promote numerical configuration	limited panel	promoted
EXP-032	product closeout	freeze closing panel	valid	promoted closeout

D.2 Rejected and non-promoted methods

Table D2. Main negative results and the information retained from them.

Method or hypothesis	Result	What remains useful
Initial eSSVI comparison	unfair repair state for promotion	motivated repair-matched rerun
Repair-matched eSSVI global replacement	failed across families	monthly-PM local candidate retained
Joint SVI	no stable broad advantage	coherent research comparator
Direct-price SVI	lower local residuals, failed shape or tail gates	established objective mismatch
Raw volatility-matching repo	noise sensitive and adverse	motivated smoothing and exact-support design
Smoothed local-window repo	80 local wins, adverse aggregate	family-level hypotheses retained
Full American IV normalization	localized and mixed	feasible research implementation
Premium-adjusted normalization	no new promotion context	targeted diagnostic use
Strike-side influence cap	local success, poor transfer	diagnosed concentrated influence
Boundary-primary hypothesis	low error mass at endpoints	redirected work to interior influence
Strong price-aware penalties	unstable in adverse slices	justified weak guarded treatment
Common-support weighting	did not resolve target	support mismatch not dominant
Residual local weighting	little effect on main target	guard behavior confirmed
Dupire local volatility as repair	not implemented	retained only for a future dynamics objective

Appendix E. Convergence Evidence

E.1 CRR ladder

Table E1. Absolute difference from the 3,200-step CRR family reference.

Steps	Near-ATM call	OTM call
100	0.000075	0.000842
200	0.001346	0.000497
400	0.001380	0.001373
800	0.000971	0.000350
1,600	0.000466	0.000277
3,200	reference	reference

The nonmonotone sequence is retained. Reporting only the 100- and 3,200-step values would conceal the lattice oscillation relevant to the 400-step policy.

E.2 CN-PSOR ladder

Table E2. Absolute difference from CRR reference by PDE configuration.

Time/space grid	Spot scale	Near-ATM call	OTM call
100 by 100	5	2.049884	0.384657
200 by 200	5	1.157015	0.153250
400 by 400	5	0.322795	0.048802
800 by 800	5	0.058965	0.008362
1,200 by 1,200	5	0.021625	0.004366
1,600 by 1,600	3	0.001571	0.001828
1,600 by 1,600	4	0.005241	0.000605
1,600 by 1,600	5	0.010050	0.003032
1,600 by 1,600	6	0.023638	0.006996
1,600 by 1,600	7	0.035824	0.009964

The scale-3 grid is closer to CRR for the near-ATM contract, while scale 4 is closer for the OTM contract. Scale 4 was selected as a balanced independent configuration, not because it minimizes every observed contract error. The promoted 1,600-by-1,600 scale-4 row differs from CRR by 0.005241 and 0.000605; cross-family difference is not expected to vanish exactly at finite resolution.

Appendix F. Per-Asset Results

F.1 Pricing and quality

Table F1. Per-asset closing quality.

Asset	Rows	Combined pass	Combined warn	Combined fail	Quote pass	Mean relative spread	Mean engine difference
INTC	400	10.75%	88.75%	0.50%	86.00%	0.0854	0.00032
NDX	400	0.00%	100.00%	0.00%	48.50%	0.0542	0.09148
NVDA	400	43.00%	57.00%	0.00%	83.00%	0.0844	0.00121
RUT	400	26.00%	73.00%	1.00%	70.50%	0.0680	0.01157
SPX	400	0.00%	99.75%	0.25%	62.75%	0.0433	0.01621
SPY	400	0.00%	100.00%	0.00%	92.75%	0.0375	0.00247

Combined warnings include surface and independent-engine diagnostics, so a zero combined pass rate does not imply invalid prices. NDX's mean engine difference is larger in raw dollars than the equity values and remains below the configured failure rules at the row level.

F.2 Early-exercise premium tails

Table F2. American premium distribution by asset and option side.

Asset	Side	Pairs	Mean	Median	P90	P99	Maximum
INTC	Call	260	0.000000	0.000000	0.000000	0.000000	0.000000
INTC	Put	140	0.001411	0.000761	0.003468	0.009428	0.009912
NVDA	Call	170	0.000014	0.000000	0.000000	0.000383	0.000474
NVDA	Put	230	0.005388	0.000360	0.009363	0.046089	0.417179
SPY	Call	94	0.000149	0.000000	0.000009	0.004731	0.007657
SPY	Put	306	0.011974	0.001840	0.031136	0.132532	0.289669

Appendix G. Reproduction Commands

All commands run from the extracted public repository root.

G.1 Python environment and tests

```
python -m venv .venv
.\.venv\Scripts\python.exe -m pip install --upgrade pip
.\.venv\Scripts\python.exe -m pip install -e ".[dev,report]"
.\.venv\Scripts\python.exe -m ruff check . --no-cache
.\.venv\Scripts\python.exe -m pytest -p no:cacheprovider
```

G.2 Synthetic pipeline

```
.\.venv\Scripts\python.exe examples\generate_synthetic_chain.py --output build\synthetic_chain.csv
.\.venv\Scripts\volsurf.exe snapshot --input build\synthetic_chain.csv --output build\snapshot
.\.venv\Scripts\volsurf.exe benchmark --input build\synthetic_chain.csv --output build\benchmark
.\.venv\Scripts\volsurf.exe validate --bundle build\benchmark
.\.venv\Scripts\volsurf.exe evidence --output build\evidence
```

G.3 Public evidence

```
.\.venv\Scripts\python.exe tools\build_report_evidence.py
.\.venv\Scripts\python.exe tools\validate_report_evidence.py
```

The maintainer-only curator requires an explicit private archive root and is not part of public reproduction. Its output is the 20 sanitized tables already packaged under `evidence/data`.

G.4 Technical report

```
.\.venv\Scripts\python.exe tools\build_technical_report.py
.\.venv\Scripts\python.exe tools\validate_technical_report.py
```

The build produces Markdown, HTML, DOCX, and PDF outputs and records their hashes in `report/report_manifest.yaml`. Rendering and page-image review are separate final acceptance steps.

G.5 C++ library

```
cmake -S cpp -B build\cpp -DCMAKE_BUILD_TYPE=Release
cmake --build build\cpp --config Release
ctest --test-dir build\cpp -C Release --output-on-failure
cmake --install build\cpp --config Release --prefix build\install
```

Appendix H. Public and Private Data Boundary

H.1 Public artifacts

The public evidence includes context aggregates, sufficient statistics, synthetic explanations, claim checks, source hashes, and figures. It excludes bid, ask, strike, spot, raw market price, model price, option identifier, contract key, and private path fields. The exact-support repo join occurs inside the private curation stage; only counts and error sums cross the boundary.

The source manifest contains 838 opaque records with role, byte count, and hash. The identifiers are not vendor contract identifiers and cannot be used to reconstruct a chain. The 20 public empirical tables contain only the granularity required for the published claims.

H.2 Reproducibility claim

Public empirical claims are reproducible from packaged derived data. Source-chain reconstruction is not public and is not claimed. Mathematical identities and synthetic workflows are fully reproducible from original code and generated inputs. The two frozen private test-count claims remain documented-only; current public software claims use the public release evidence.

H.3 License

The MIT license applies to original source code and report-production code. It does not relicense third-party market data, publications, or dependencies. Aggregate evidence is provided for research traceability without conveying rights to the source data.

Appendix I. Claim and Source Traceability

I.1 Claim eligibility

Table I1. Public claim status. Full statements, values, units, evidence files, recomputation methods, and caveats are in `evidence/claim_ledger_public.csv`.

Claim	Category	Public use	Verification
CLM-DAT-001	data	context and report	pass
CLM-DAT-002	data	context only	pass
CLM-DAT-003	data	report	pass
CLM-MTH-001	mathematics	report and docs	pass
CLM-MTH-002	mathematics	report and docs	pass
CLM-MTH-003	mathematics	report and docs	pass
CLM-MTH-004	mathematics	report and docs	pass
CLM-MTH-005	mathematics	report and docs	pass
CLM-TAX-001	taxonomy	report and docs	pass
CLM-SRF-001	surface research	report	pass
CLM-SRF-002	surface research	report and figure	pass
CLM-SRF-003	surface research	report and figure	pass
CLM-SRF-004	surface research	report and figure	pass
CLM-SRF-005	surface research	report	pass
CLM-SRF-006	surface research	report	pass
CLM-SRF-007	surface research	report	pass
CLM-AUD-001	endogeneity audit	report and table	pass
CLM-AUD-002	endogeneity audit	report and table	pass
CLM-AUD-003	endogeneity audit	report	pass
CLM-AUD-004	endogeneity audit	report and table	pass
CLM-BND-001	boundary audit	report	pass
CLM-AMR-001	American research	report	pass
CLM-NUM-001	numerics	report and figure	pass
CLM-NUM-002	numerics	report and figure	pass
CLM-CLS-001	closing panel	report and table	pass
CLM-CLS-002	closing panel	report and table	pass
CLM-CLS-003	closing panel	report and table	pass
CLM-CLS-004	closing panel	report	pass
CLM-CLS-005	closing panel	report and table	pass
CLM-CLS-006	closing panel	report	pass
CLM-CLS-007	closing panel	report	pass
CLM-TST-001	validation	context only for current report	documented only
CLM-TST-002	validation	context only for current report	documented only
CLM-REP-001	repo research	report and figure	pass
CLM-REL-001	validation	report and docs	pass

The current report does not use CLM-TST-001 or CLM-TST-002 to state public release quality. They describe the pre-restructure private freeze and are retained for historical context.

I.2 Source registry

Table I2. Primary source identifiers. Full citations and verification metadata are in `report/source_registry.csv`.

Source	Citation	Project use
SRC-001	Stoll (1969)	European parity
SRC-002	Black and Scholes (1973)	European valuation and PDE
SRC-003	Merton (1973)	option-pricing foundations and dividends
SRC-004	Black (1976)	forward-form pricing
SRC-005	Brent (1971)	bracketed roots
SRC-006	Fritsch and Carlson (1980)	monotone curve interpolation
SRC-007	Ayer et al. (1955)	isotonic projection
SRC-008	Breeden and Litzenberger (1978)	strike convexity
SRC-009	Gatheral and Jacquier (2014)	SVI and static arbitrage
SRC-010	Hendriks and Martini (2019)	eSSVI
SRC-011	Dupire (1994)	local-volatility alternative
SRC-012	Byrd et al. (1995)	L-BFGS-B
SRC-013	Kraft (1988)	SLSQP
SRC-014	Huber (1964)	robust loss
SRC-015	Fan (1992)	local-linear smoothing
SRC-016	Efron (1979)	paired bootstrap context
SRC-017	Cox, Ross, and Rubinstein (1979)	binomial pricing
SRC-018	Leisen and Reimer (1996)	smooth tree convergence
SRC-019	Crank and Nicolson (1947)	theta time stepping
SRC-020	Rannacher (1984)	startup damping
SRC-021	Cryer (1971)	projected over-relaxation
SRC-022	Brennan and Schwartz (1978)	finite-difference option pricing
SRC-023	Kim (1990)	early-exercise-premium representation
SRC-024	Carr, Jarrow, and Myneni (1992)	American put decomposition
SRC-025	Whaley (1981)	known-dividend American calls

I.3 Traceability files

The authoritative machine-readable traceability set is:

- `evidence/claim_ledger_public.csv` for empirical and implementation claims;
- `evidence/claims_recomputed.csv` for 66 independent claim components;
- `evidence/figure_manifest.csv` for figure inputs, scope, units, and caveats;
- `report/equation_traceability.csv` for equations and code symbols;
- `report/source_registry.csv` for primary literature verification;
- `experiment_registry.csv` for research decisions and invalidation state.