WHEN SAME SIGN IS NOT ENOUGH: AUDITING MODALITY-DOMINANCE CLAIMS

Anonymous authors
Paper under double-blind review

ABSTRACT

A modality score is not yet a modality-dominance claim. Prior work established that removal semantics can change explanations and that importance estimates can be uncertain, but these issues are usually treated separately. We ask whether a directional modality claim remains statistically resolved across a declared family of plausible interventions. We audit a fixed text–numeric predictor under three removal operators, separate the observed reliance gap from an operator-specific empirical no-signal reference, and use paired inference that propagates dependent-sample and finite-null-bank uncertainty. In a leakage-audited financial forecasting stress test, all 21 operator–measurement null-centred point estimates favour numeric reliance. Yet only $5/7$ measurements are resolved across the full operator family; $2/7$ remain unresolved under same-date regime-matched removal. A prespecified marginal-versus-regime-matched magnitude contrast excludes zero on all seven measurements. Thus complete point-sign agreement can coexist with incomplete family-level inferential resolution. The contribution is a claim-level audit that determines when intervention semantics and uncertainty support a modality-dominance statement; the underlying attribution score is unchanged.

1 INTRODUCTION

Multimodal studies routinely turn attribution or ablation scores into statements such as "modality A dominates modality B." That step is scientifically stronger than reporting a score. A dominance claim asserts that one input stream matters more under a particular notion of removal, relative to an implicit or explicit reference, and with enough statistical support to justify a directional conclusion. If any of those ingredients changes, a stable-looking ranking can overstate what the evidence actually resolves.

Two established literatures expose the ingredients but not this claim-level question. Removal-based explanation shows that "feature absent" is not unique: marginal, conditional, baseline, and interventional constructions can answer different questions under dependence (Janzing et al., 2020; Aas et al., 2021; Covert et al., 2021; Martin and Haufe, 2026). Uncertainty-aware explanation shows that a point attribution or ranking need not justify a definite scientific ordering (Williamson and Feng, 2020; Marx et al., 2023; Laberge et al., 2023). We do *not* claim either observation as new. The missing scientific question is their conjunction: *if a paper reports that one modality dominates another, does that conclusion remain statistically resolved over a declared family of scientifically plausible interventions?*

We study modality dominance as an *operator-indexed inferential claim*. For each removal operator, a fixed confidence-loss functional produces a text–numeric reliance gap D. Because the complete intervention pipeline can itself have an operator-dependent no-signal offset, we keep D distinct from a null-centred response $D^* = D - \mu_0$. A paired bootstrap then propagates dependent evaluation-sample and finite-null-bank uncertainty, with simultaneous intervals over the three declared operators within each measurement. We use these established components to audit the directional claim itself.

The empirical result contains no ranking flip. Across seven prespecified measurements and three prespecified operators, all 21 D^* point estimates have the same negative sign. Nevertheless, only $5/7$ measurements are family-resolved; $2/7$ remain unresolved because the same-date regime-matched

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

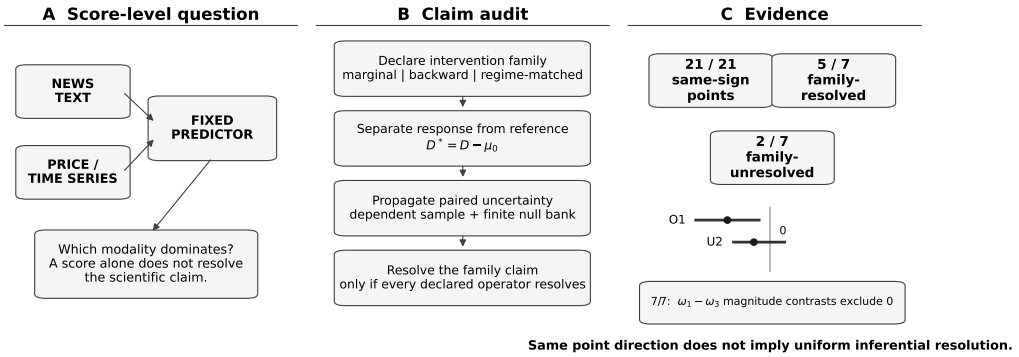

Figure 1: Claim-audit overview. (A) A fixed text–numeric predictor yields a modality-ranking question. (B) The audit declares removal semantics, separates observed response from its empirical reference, and propagates paired uncertainty before assigning a family-level direction. (C) The reported panel has $21/21$ same-sign point estimates, $5/7$ family-resolved measurements, $2/7$ unresolved measurements, and nonzero ω_1–ω_3 magnitude contrasts on $7/7$ measurements; O1 and U2 illustrate resolved and unresolved simultaneous intervals.

interval includes zero. The prespecified marginal-versus-regime-matched magnitude contrast excludes zero on $7/7$ measurements. Figure 1 summarizes this separation. Point direction can be completely stable while the strength of the scientific conclusion remains operator-dependent.

Contributions. We make the reported modality-dominance statement, rather than an individual attribution score, the object of inference by separating raw response, empirical reference, operator-specific uncertainty, and family-level resolution. Empirically, we show that complete point-sign agreement can coexist with incomplete resolution over the declared operator family. We also separate confirmatory direction, operator-magnitude contrasts, negative controls, and diagnostics so that secondary analyses do not alter the primary claim. Finance provides a dependent, regime-changing text–numeric stress test; the same audit applies whenever modality dominance is inferred through masking, replacement, imputation, or perturbation rather than observed directly.

2 Related Work

Removal semantics. Removal-based explanations differ in *how* information is made absent, *what* model behaviour is queried, and *how* influence is summarized (Covert et al., 2021). Under dependent inputs, marginal, conditional, interventional, and baseline constructions need not represent the same scientific question (Janzing et al., 2020; Aas et al., 2021; Sundararajan and Najmi, 2020; Ren et al., 2023). Recent work improves estimation or makes support/causal-context assumptions more explicit, but does not eliminate this semantic choice (Jethani et al., 2022; Chen et al., 2023; Enouen and Liu, 2025; Bhattacharjee et al., 2025; Martin and Haufe, 2026). Accordingly, our contribution is *not* the observation that perturbation or reference choice matters.

Uncertainty in explanations. Population feature-importance inference and uncertainty-aware explanation quantify uncertainty arising from sampling, an unknown optimal model, Rashomon sets, decision-boundary geometry, or explanation approximation (Williamson and Feng, 2020; Marx et al., 2023; Laberge et al., 2023; Hill et al., 2024; Alkhatib et al., 2024). These works ask when a feature-importance or explanation object is trustworthy. Our uncertainty axis is different: the fitted predictor and measurement panel are held fixed while the *intervention-defined estimand itself* varies over a declared operator family, with dependent-sample and finite-reference uncertainty propagated for the resulting directional claim.

Modality imbalance in multimodal learning. MM-SHAP measures relative modality contribution with Shapley values (Parcalabescu and Frank, 2023); sample-level modality valuation reveals heterogeneous contribution imbalance across examples (Wei et al., 2024); and optimization/theory

work studies or mitigates unimodal bias and modality competition (Peng et al., 2022; Zhang et al., 2024; Hua et al., 2024; Wang et al., 2025; Gong et al., 2025). Their primary object is a contribution score, a training pathology, or a rebalancing method. Our primary object is instead the *evidentiary status of a reported directional statement*: after several plausible intervention semantics are declared, which "modality A dominates modality B" conclusions remain statistically resolved?

Positioning. These ingredients already exist separately: removal semantics, attribution/importance uncertainty, and modality-contribution analysis. We combine them at the level of the reported directional claim. The empirical contribution is the observed separation among *sign stability*, *magnitude sensitivity*, and *inferential resolution* over a prespecified operator family. The paired bootstrap, null bank, and confidence-loss functional are supporting machinery rather than novelty claims.

Financial setting. FNSPID and recent forecasting work combine financial text with structured time-series inputs, and related finance systems use attribution to interpret such models (Dong et al., 2024; Koval et al., 2024; Huang et al., 2025; Kim and Park, 2025). We use this setting because temporal dependence, regime variation, and heterogeneous representations make intervention assumptions consequential. The transferable lesson concerns how multimodal dominance claims are justified; the observed numeric-over-text ordering remains specific to this financial setting.

3 Claim Audit

3.1 Operator-indexed intervention response

We fix a probabilistic predictor $p_\theta(y \mid x)$ with text x_t and numeric x_n representation branches. The declared family

$$\Omega = \{\omega_1, \omega_2, \omega_3\}$$

contains marginal permutation, backward-within-ticker donation, and same-date regime-matched donation. They deliberately preserve increasing amounts of contextual structure: an empirical branch marginal, ticker-specific historical context, and contemporaneous regime context, respectively. The third operator is *not* a conditional-replacement estimator. For branch j, donor $d_{i,j}^{(\omega,s)}$, dose $f \in \{0, .2, .4, .6, .8, 1\}$, and fixed corruption seed s, the published corruption path is

$$\widetilde{x}_{i,j}^{(f,\omega,s)} = \sqrt{1-f}\, x_{i,j} + \sqrt{f}\, d_{i,j}^{(\omega,s)}. \tag{1}$$

Let $\widehat{y}_i = \arg\max_y p_\theta(y \mid x_i)$ be the original predicted class and $c_i(x) = p_\theta(\widehat{y}_i \mid x)$. The branch response curve keeps this class fixed:

$$g_{m,j,\omega}(f) = \frac{1}{N_m} \sum_{i=1}^{N_m} \left[c_i(x_i) - \mathbb{E}_s\, c_i\Big(x_{i,-j}, \widetilde{x}_{i,j}^{(f,\omega,s)}\Big) \right]. \tag{2}$$

Trapezoidal integration over the dose grid gives $R_{m,j,\omega} = \mathrm{Trapz}_f\ g_{m,j,\omega}(f)$. This definition matters because it fixes both the prediction being scored and the perturbation trajectory; a different donor family or path defines a different operational question. The raw directional gap is

$$D_{m,\omega} = R_{m,t,\omega} - R_{m,n,\omega}. \tag{3}$$

Negative D means greater numeric reliance under *that* path and operator. It is an operational response of a fixed fitted model, not a structural-causal effect or an operator-invariant property.

The path itself can change scale. With $q_\omega = \mathbb{E}\|d^{(\omega)}\|^2/\mathbb{E}\|x\|^2$ and $\rho_\omega = \mathbb{E}\langle d^{(\omega)}, x\rangle/\mathbb{E}\|x\|^2$,

$$\frac{\mathbb{E}\|\widetilde{x}^{(f,\omega)}\|^2}{\mathbb{E}\|x\|^2} = (1-f) + f q_\omega + 2\sqrt{f(1-f)}\,\rho_\omega. \tag{4}$$

Thus scalar second moment is preserved for every dose when $q_\omega = 1$ and $\rho_\omega = 0$; independence, equal covariance, and pointwise equal donor norm are stronger, non-necessary conditions.

3.2 EMPIRICAL REFERENCE AND PAIRED INFERENCE

Zero need not be the centre produced by the full pipeline after predictive label information is removed. For each measurement and operator, let $D^{(r)}_{0,m,\omega}$ be the response of no-signal reference model r, with $K = 50$ models in the full bank. We keep the raw estimand distinct from

$$\mu_{0,m,\omega} = \frac{1}{K}\sum_{r=1}^{K} D^{(r)}_{0,m,\omega}, \qquad D^{*}_{m,\omega} = D_{m,\omega} - \mu_{0,m,\omega}. \tag{5}$$

The null generator is a within-ticker temporal-block construction; it does not preserve same-date cross-sectional dependence. Accordingly, μ_0 is a procedure-specific empirical noise reference rather than a finite-sample-valid equal-reliance null.

We use $B = 1000$ circular block-bootstrap draws (Politis and Romano, 1992). Within draw b, all three operators use the same resampled rows, fixed donor assignments, and corruption seeds. We additionally sample 50 null-model indices with replacement from the full bank and compute

$$D^{*,(b)}_{m,\omega} = D^{(b)}_{m,\omega} - \frac{1}{50}\sum_{k=1}^{50} D^{(b,r_k^{(b)})}_{0,m,\omega}. \tag{6}$$

This propagates evaluation-sample and finite-null-bank Monte Carlo uncertainty while retaining covariance among operator estimates. It does not propagate retraining, target selection, alternative null generators, or operator families.

For each measurement, let $\widehat{\sigma}_{m,\omega}$ be the bootstrap standard deviation and form

$$T^{(b)}_{m} = \max_{\omega\in\Omega}\left|\frac{D^{*,(b)}_{m,\omega} - \overline{D}^{*}_{m,\omega}}{\widehat{\sigma}_{m,\omega}}\right|.$$

The empirical 95th percentile of $\{T^{(b)}_m\}_{b=1}^{B}$ produces sup-t intervals $C^{\mathrm{sim}}_{m,\omega} = [L_{m,\omega}, U_{m,\omega}]$. The max is over three operators *within* one measurement only; no family-wise statement is made across the seven overlapping measurements. This distinction is important because the scientific claim is indexed by intervention semantics, while the seven measurements are heterogeneous confirmations rather than i.i.d. replicates.

We separate three levels of evidence. First, $\mathrm{sign}(D^{*}_{m,\omega})$ records point-direction agreement. Second, an individual operator is resolved numeric when $U_{m,\omega} < 0$ (or resolved text when $L_{m,\omega} > 0$); an interval containing zero is unresolved, not evidence of equality. Third, the measurement is *family-resolved numeric* only if

$$\max_{\omega\in\Omega} U_{m,\omega} < 0. \tag{7}$$

Otherwise it is family-unresolved; unresolved does not imply equality. The prespecified secondary contrast $\Delta^{*}_{m,1,3} = D^{*}_{m,\omega_1} - D^{*}_{m,\omega_3}$ is computed draw by draw from the same paired bootstrap. Its interval answers whether two operators assign different *magnitudes*; it does not determine the sign of D^{*}_{m,ω_3} and cannot upgrade an unresolved measurement.

For each measurement we report raw D, the operator-specific full-bank centre μ_0, centred D^{*}, simultaneous operator intervals, the family-level status, and the secondary ω_1–ω_3 contrast. Keeping these objects separate prevents three common logical substitutions: point-sign agreement for statistical resolution, nonzero operator contrast for dominance under the weaker operator, and empirical null centring for a universal no-signal definition. Appendix A.1 records seeds and the exact uncertainty scope.

4 EXPERIMENTAL DESIGN

Data, probe, and leakage controls. We use FNSPID (Dong et al., 2024) on U.S. equities from 2015–2023. Text contains only news timestamped at or before each reference date, encoded with `all-MiniLM-L6-v2` (a Sentence-BERT-family encoder (Reimers and Gurevych, 2019)) and reduced by train-fitted PCA to 50 components; Appendix A.3 gives the exact text window, truncation

rule, indicator list, target definition, and probe settings. The numeric branch summarizes a 64-trading-day window with 63 log-return coordinates and eight technical indicators. Standardization and PCA are fit on training data only. A fixed L_2-regularized logistic probe is evaluated chronologically with a 10-trading-day stride and a 20-trading-day embargo. Automated checks enforce no train/test row overlap and no post-reference-date news.

The analysis panel comprises three original-universe slices (train/test $860/901$, $1761/804$, $2573/689$), three slices from a disjoint 19-ticker universe ($969/1007$, $1976/1007$, $2983/1007$), and one later 2023-H2 temporal confirmation ($3262/169$). The two universes are disjoint by ticker, while T1 shifts the evaluation period. Expanding training windows overlap, however, so the seven measurements are not independent replications and are never pooled into a cross-measurement family-wise guarantee. This design intentionally provides several forms of confirmation without pretending they are seven exchangeable trials.

The probe is deliberately simple. We do not ask whether a new architecture can exploit text–numeric interactions better; we ask whether a scientific reliance statement about a *fixed fitted predictor* survives reasonable intervention semantics. Holding the predictor fixed prevents retraining variability from being mixed with the operator question, while also limiting the scope of the conclusion to this fitted model class.

Predictive-feasibility gate. Before reliance analysis, a prespecified menu screened four targets: 3-day direction, 3-day abnormal return versus QQQ, 3-day cross-sectional rank, and next-10-day realized-volatility regime. Only volatility regime passed the gate (mean joint AUC 0.582); Appendix A.3 reports the full screen. The gate is a feasibility prerequisite, not evidence for modality dominance: a reliance audit is uninformative if the joint predictor itself carries no useful signal. Because target choice used the original-universe panel, O1–O3 retain selection dependence. U1–U3 and T1 were not used for that choice and therefore provide confirmation outside the selection panel, although the reported intervals do not formally propagate the earlier target-selection uncertainty.

Claim–evidence map. Table 1 makes the evidentiary hierarchy explicit before the Results are read.

Table 1: Claim–evidence map. Only the first row determines the family-level labels; the remaining rows answer distinct secondary or explanatory questions.

Object	Direct evidence	Result	Role
Family-level direction	7 measurements $\times$ 3 operators; paired simultaneous D^* intervals	$21/21$ same-sign points; $5/7$ resolved, $2/7$ unresolved	Confirmatory
Operator magnitude	Paired OP1–OP3 Δ^*, all 7 measurements	Contrast excludes zero on $7/7$	Secondary
Scalar-path explanation	Exact second-moment normalization, 2 diagnostic cases	Gap remains nonzero on both	Falsifier
Sign-reversal mechanism	Exact marginal vs. conditional comparison	No reversal; both gaps favour text	Negative
Geometry diagnostic	Decision-axis displacement, 2 diagnostic cases	Strong association in U2; weaker in O1	Diagnostic

For the primary paired analysis, the three-operator family, five corruption seeds, full 50-model null bank, $B = 1000$ paired resampling plan, family-resolution rule, and ω_1–ω_3 secondary contrast were prespecified before the uniform panel statistics were computed. The hierarchy is deliberately asymmetric. RQ1 alone determines whether a measurement is family-resolved; RQ2 quantifies a distinct magnitude hypothesis; RQ3 and RQ4 can only remove candidate explanations; RQ5 provides explanatory support after the primary analysis is fixed. This prevents a later diagnostic from redefining the confirmatory endpoint. The reported mismatch and Shapley-reference analyses do not support comparative reliance-functional superiority, operator invariance, or a unique mechanism.

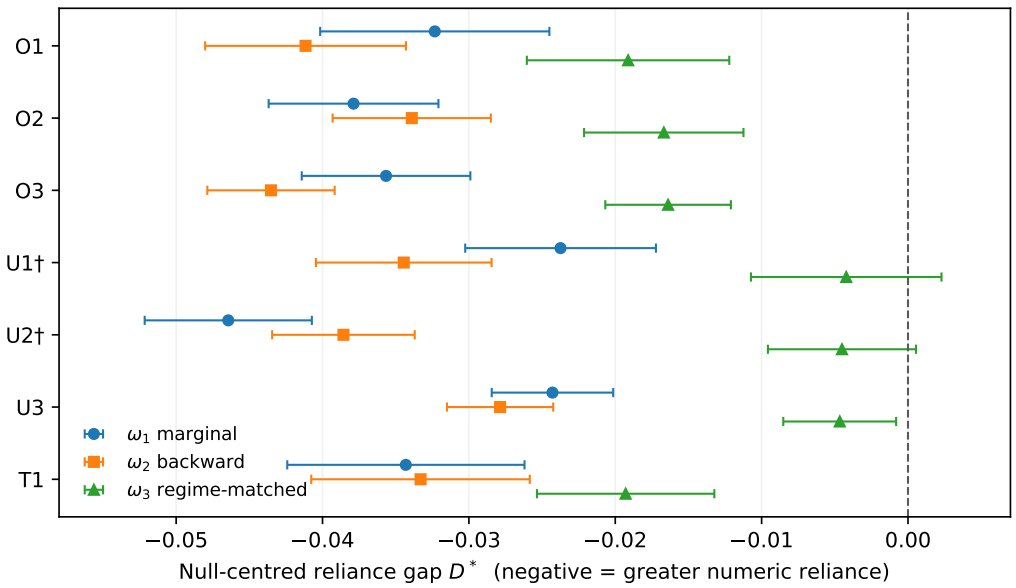

Figure 2: Operator-family estimates. All 21 D^* point estimates are negative, but U1† and U2† remain family-unresolved because the regime-matched interval crosses zero; the other five measurements are family-resolved numeric. Intervals are simultaneous across operators within each measurement, not across measurements.

5 Results

RQ1 tests family-level direction, RQ2 tests operator-magnitude differences, and RQ3–RQ5 evaluate candidate explanations without changing the primary label.

5.1 RQ1: Complete sign agreement, incomplete claim resolution

Figure 2 visualizes the primary paired panel and Appendix A.2 reports exact values. All 21 empirical-null-centred point estimates are negative. At the level of point estimates alone, numeric reliance therefore appears completely stable across the tested intervention family.

The inferential conclusion is weaker. Only five measurements have all three simultaneous operator intervals below zero. U1 and U2 remain family-unresolved because the same-date regime-matched operator crosses zero, with OP3 intervals $[-0.01073, +0.00229]$ and $[-0.00957, +0.00055]$. U3, from the same disjoint universe, is resolved, so the two unresolved cases do not support a universe-level pattern. Both unresolved cases also lie among U1–U3/T1, the four confirmation measurements not used to choose the target; the failure of family-level resolution is therefore not confined to the target-selection panel.

The unresolved status is localized to OP3: OP1 and OP2 are strictly negative on all seven measurements. We report this descriptively rather than ranking operators by validity, because the three constructions encode different removal semantics. The empirical reference also matters independently of operator choice. Under ω_1, the full-bank null centre is negative on all seven measurements (-0.02508 to -0.00166). Raw D and centred D^* therefore answer different questions: D records the fitted model's response to a perturbation, while D^* asks how much of that direction exceeds the offset produced by the same pipeline under the prespecified no-label-signal construction.

The primary failure mode is the following:

$$\boxed{21/21 \text{ same-sign point estimates } \not\Rightarrow \ 7/7 \text{ resolved dominance claims.}}$$

The observed conclusion is $5/7$ family-resolved and $2/7$ family-unresolved.

5.2 RQ2: Operator semantics alter magnitude without changing sign

The prespecified paired contrast

$$\Delta^*_{m,1,3} = D^*_{m,\omega_1} - D^*_{m,\omega_3}$$

excludes zero on all seven measurements. Point contrasts range from -0.01320 to -0.04193: after empirical centring, marginal permutation yields a more negative text–numeric gap than same-date regime-matched donation throughout the reported panel.

This contrast answers a different question from RQ1. In U1 and U2, OP1–OP3 clearly differ in magnitude while the OP3 sign itself remains unresolved. Thus

"the operators differ" $\neq$ "the same direction is resolved under every operator."

The secondary contrast quantifies dependence on donor semantics; it does not rescue a weaker primary sign.

5.3 RQ3–RQ5: Negative controls narrow explanations; geometry remains diagnostic

We next ask what might explain the ω_1–ω_3 magnitude gap. These analyses do not alter the prespecified RQ1 labels.

Scalar path energy is not sufficient in the two diagnostic cases. Exact dose-wise scalar second-moment normalization leaves the contrast nearly unchanged. U2 moves from $-0.05320[-0.05863, -0.04789]$ to $-0.05248[-0.05757, -0.04730]$; O1 moves from $-0.03017[-0.03635, -0.02373]$ to $-0.02938[-0.03532, -0.02310]$. All intervals exclude zero. This falsifies only the simple explanation that gross scalar second-moment distortion is sufficient. Featurewise scale, covariance, support, angular, and model-direction effects remain open.

The analytic sign-reversal mechanism does not transfer to confidence loss. In the controlled $\sigma = 0$ comparison, marginal $D = +0.01270[+0.01174, +0.01366]$ and exact-conditional $D = +0.04118[+0.04066, +0.04168]$ both favour text. The squared-change sign-reversal example therefore does not justify a generic sign-nonidentifiability claim for the fitted confidence-loss endpoint.

Decision-axis geometry remains diagnostic. In diagnostic case U2, the OP3/OP1 numeric displacement ratio is $0.578[0.537, 0.621]$, while the text control is $0.975[0.946, 1.005]$. On O1, the numeric ratio moves in the same direction but is weaker, $0.792[0.759, 0.826]$, with text $1.025[0.988, 1.061]$. The secondary numeric interval crosses the prespecified 0.8 strong-effect threshold. The result is therefore consistent with model-relative geometry contributing to the observed magnitude gap on the tested cases, but it neither identifies a necessary/causal mechanism nor changes the primary panel labels.

Across the declared family, *sign stability, magnitude sensitivity, and inferential resolution are empirically distinct*. This separation is the central scientific finding; the paper does not propose a new attribution score or rank the operators by validity.

6 Discussion

Claim-level consequence of intervention sensitivity. Known removal-semantics work already implies that different interventions can change an attribution value. The claim-level consequence is stronger: *complete point-sign agreement can still fail to support a uniformly resolved directional statement*. The observed separation identifies an inferential failure mode in which a ranking looks qualitatively stable while the strength of the dominance claim changes.

Measurement design and claim design should be separated. Measurement design chooses a reliance functional, intervention construction, and empirical reference. Claim design specifies what must remain true before those measurements can be summarized as "modality A dominates modality B." Conflating the two lets one score carry semantics and inference that it was never designed

to encode. Our audit makes the separation explicit: raw response and empirical reference are distinct; operator-specific intervals are distinct from point signs; and the family label is an intersection statement over declared operators. Secondary magnitude contrasts answer a different hypothesis and cannot substitute for unresolved primary evidence.

Role of negative controls. The two negative controls and one diagnostic prevent the empirical operator gap from being converted into an unsupported mechanism claim. Scalar second-moment normalization rules out one simple artefact in two diagnostic cases; the controlled comparison rejects a stronger sign-nonidentifiability interpretation; and decision-axis geometry remains diagnostic because the second case is weaker. Together, these analyses specify which interpretations the primary result *does not* support.

Scope beyond finance. The transferable result is a reporting discipline for intervention-based modality dominance; the observed numeric-over-text ordering remains specific to this market dataset. Whenever dominance is inferred through masking, replacement, imputation, ablation, or donor-based perturbation, a defensible claim should identify the intervention family, state any procedure-specific reference, report operator-specific uncertainty, separate magnitude sensitivity from sign resolution, and permit an unresolved outcome. Finance is useful here because dependence, regime variation, and heterogeneous modalities stress these assumptions; the audit logic does not depend on financial semantics.

Reporting checklist. A defensible multimodal dominance claim should state: *what was removed, relative to what reference, with what uncertainty, over which declared operator family, and whether every operator resolved the same direction*. This makes sensitivity visible without equating it with failure and preserves the scientifically valid outcome that the claim remains unresolved.

7 Limitations and Scope

The audit conditions on one fitted L_2-regularized logistic probe, one adaptively selected volatility-regime target, one U.S.-equity text–numeric pipeline, and three donor constructions. Target selection used O1–O3; U1–U3 and T1 reduce but do not remove selection dependence. Family-resolved means resolved over this declared operator family, not invariant to every possible removal scheme, architecture, training seed, or representation learner.

The within-ticker temporal-block null is a procedure-specific empirical noise reference and does not preserve same-date cross-sectional dependence. The bootstrap propagates evaluation-sample and finite-null-bank uncertainty, not retraining, target selection, alternative null generators, or alternative operator families. Simultaneous intervals cover three operators within each measurement, not the seven overlapping measurements jointly. Two scale and two geometry diagnostic cases cannot establish a general mechanism. Unresolved does not imply equality; resolved does not imply causality. Broader external validity requires new domains, architectures, interventions, nulls, and independently designed measurements. A particularly important next test is whether the $21/21$-versus-$5/7$ separation also appears in richer nonlinear multimodal models, where removal may induce larger off-manifold shifts and model retraining uncertainty may no longer be negligible. The present paper establishes the audit logic and one stress-test instance; its prevalence across multimodal ML remains unknown.

8 Conclusion

A stable modality score is not yet a defensible modality-dominance claim. Across the declared operator family, all 21 null-centred point estimates favour numeric reliance, yet only five of seven measurements remain resolved across all three operators. Modality-dominance statements should therefore make intervention semantics, reference construction, uncertainty, and unresolved cases explicit before being treated as properties of a model.

Reproducibility Statement

The main text specifies the fixed predictor, operator family, estimands, paired inference, measurement panel, and evidence roles. Appendix A.1 records the corruption seeds, 50-model null bank, path identity, uncertainty scope, and reported quantities; Appendix A.2 gives every primary paired-panel estimate, simultaneous interval, family label, and operator contrast; Appendix A.3 records measurement sizes and target-gate results. An anonymised supplementary archive accompanies this submission: it contains, for every round reported here, the frozen specification, the run script, the verification tests, and the machine-readable results, together with the errata recording withdrawn claims. Each round re-derives the frozen marginal-operator estimates as a reproduction control before any new quantity is interpreted. All local filesystem paths are replaced by placeholders, so the scripts require a data directory to be supplied. The financial source data (FNSPID) is public and is not redistributed.

AI Use Statement

AI tools were employed for limited assistance in language refinement and in carrying out experiments based on procedures and settings established by the authors. The formulation of the research problem, hypotheses, methods, experimental setup, analysis of results, and scientific conclusions remained entirely under human authorship. Any material produced with AI assistance was independently checked by the authors, who assume full responsibility for the accuracy and integrity of the submitted work.

References

Sundararajan, Mukund and Najmi, Amir. The Many Shapley Values for Model Explanation. Proceedings of the 37th International Conference on Machine Learning, 119:9269–9278, 2020.

Janzing, Dominik and Minorics, Lenon and Blöbaum, Patrick. Feature Relevance Quantification in Explainable AI: A Causal Problem. Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, 108:2907–2916, 2020.

Aas, Kjersti and Jullum, Martin and Løland, Anders. Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values. Artificial Intelligence, 298:103502, 2021.

Covert, Ian and Lundberg, Scott and Lee, Su-In. Explaining by Removing: A Unified Framework for Model Explanation. Journal of Machine Learning Research, 22:1–90, 2021.

Jethani, Neil and Sudarshan, Mukund and Covert, Ian Connick and Lee, Su-In and Ranganath, Rajesh. FastSHAP: Real-Time Shapley Value Estimation. International Conference on Learning Representations, 2022.

Ren, Jie and Zhou, Zhanpeng and Chen, Qirui and Zhang, Quanshi. Can We Faithfully Represent Absence States to Compute Shapley Values on a DNN? International Conference on Learning Representations, 2023.

Chen, Hugh and Covert, Ian C. and Lundberg, Scott M. and Lee, Su-In. Algorithms to Estimate Shapley Value Feature Attributions. Nature Machine Intelligence, 5:590–601, 2023.

Enouen, James and Liu, Yan. InstaSHAP: Interpretable Additive Models Explain Shapley Values Instantly. International Conference on Learning Representations, 2025.

Bhattacharjee, Robi and Frohnapfel, Karolin and von Luxburg, Ulrike. How to Safely Discard Features Based on Aggregate SHAP Values. Proceedings of Thirty Eighth Conference on Learning Theory, 291:280–314, 2025.

Williamson, Brian D. and Feng, Jean. Efficient Nonparametric Statistical Inference on Population Feature Importance Using Shapley Values. Proceedings of the 37th International Conference on Machine Learning, 119:10282–10291, 2020.

Marx, Charles and Park, Youngsuk and Hasson, Hilaf and Wang, Yuyang and Ermon, Stefano and Huan, Luke. But Are You Sure? An Uncertainty-Aware Perspective on Explainable AI. Proceedings of the 26th International Conference on Artificial Intelligence and Statistics, 206:7375–7391, 2023.

Laberge, Gabriel and Pequignot, Yann and Mathieu, Alexandre and Khomh, Foutse and Marchand, Mario. Partial Order in Chaos: Consensus on Feature Attributions in the Rashomon Set. Journal of Machine Learning Research, 24:1–50, 2023.

Hill, Davin and Masoomi, Aria and Torop, Max and Ghimire, Sandesh and Dy, Jennifer. Boundary-Aware Uncertainty for Feature Attribution Explainers. Proceedings of The 27th International Conference on Artificial Intelligence and Statistics, 238:55–63, 2024.

Alkhatib, Amr and Boström, Henrik and Johansson, Ulf. Estimating Quality of Approximated Shapley Values Using Conformal Prediction. Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications, 230:158–174, 2024.

Parcalabescu, Letitia and Frank, Anette. MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers):4032–4059, 2023.

Peng, Xiaokang and Wei, Yake and Deng, Andong and Wang, Dong and Hu, Di. Balanced Multimodal Learning via On-the-Fly Gradient Modulation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition:8238–8247, 2022.

Wei, Yake and Feng, Ruoxuan and Wang, Zihe and Hu, Di. Enhancing Multimodal Cooperation via Sample-level Modality Valuation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition:27338–27347, 2024.

Zhang, Yedi and Latham, Peter E. and Saxe, Andrew M. Understanding Unimodal Bias in Multimodal Deep Linear Networks. Proceedings of the 41st International Conference on Machine Learning, 235:59100–59125, 2024.

Hua, Cong and Xu, Qianqian and Bao, Shilong and Yang, Zhiyong and Huang, Qingming. ReconBoost: Boosting Can Achieve Modality Reconcilement. Proceedings of the 41st International Conference on Machine Learning, 235:19573–19597, 2024.

Wang, Jun and Cao, Fuyuan and Xue, Zhixin and Zhao, Xingwang and Liang, Jiye. CMoB: Modality Valuation via Causal Effect for Balanced Multimodal Learning. Advances in Neural Information Processing Systems, 2025.

Gong, Baoquan and Gao, Xiyuan and Zhu, Pengfei and Hu, Qinghua and Cao, Bing. Multimodal Negative Learning. Advances in Neural Information Processing Systems, 2025.

Dong, Zihan and Fan, Xinyu and Peng, Zhiyuan. FNSPID: A Comprehensive Financial News Dataset in Time Series. Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining:4918–4927, 2024.

Koval, Ross and Andrews, Nicholas and Yan, Xifeng. Financial Forecasting from Textual and Tabular Time Series. Findings of the Association for Computational Linguistics: EMNLP 2024:8289–8300, 2024.

Huang, Ruizheng and Zhang, Zhicheng and Wang, Yong. Cross-MoE: An Efficient Temporal Prediction Framework Integrating Textual Modality. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing:29927–29938, 2025.

Kim, Jinwoong and Park, Sangjin. IKNet: Interpretable Stock Price Prediction via Keyword-Guided Integration of News and Technical Indicators. Proceedings of the 6th ACM International Conference on AI in Finance:709–717, 2025.

Martin, Jörg and Haufe, Stefan. cc-Shapley: Measuring Multivariate Feature Importance Needs Causal Context. Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence, 337:4323–4352, 2026.

Politis, Dimitris N. and Romano, Joseph P. A Circular Block-Resampling Procedure for Stationary Data. Exploring the Limits of Bootstrap:263–270, 1992.

Reimers, Nils and Gurevych, Iryna. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing:3982–3992, 2019.

A Supplementary Audit Details

A.1 Protocol and reporting details

Corruption seeds are $\{13, 42, 777, 2024, 7\}$. The empirical reference uses all 50 no-signal reference models (seeds 5000–5049). The evaluation bootstrap uses circular within-ticker blocks of seven reference-index positions; because reference dates are sampled every 10 trading days, one block spans approximately 70 trading days. In each of 1000 paired bootstrap draws, evaluation rows follow this block design and 50 null-model indices are sampled with replacement. The empirical-reference models are refit after independently circularly shifting label blocks within each ticker under the prespecified temporal-block null generator; this construction is not claimed to preserve same-date cross-sectional dependence. The sup-t maximum is across the three operators within one measurement only.

The reported quantities are raw D; the operator-specific full-bank null centre μ_0; null-centred D^*; within-measurement simultaneous intervals; family resolved/unresolved status; and the prespecified ω_1–ω_3 magnitude contrast as secondary evidence. This inference propagates evaluation-sample and finite-null-bank variation, but not retraining, target selection, alternative null generators, or alternative operator families.

A.2 Primary operator-family values

Table 2: Primary paired-panel values. Entries are D^* with simultaneous 95% intervals across the three operators within each measurement. R/U denote family-resolved numeric / family-unresolved.

ID	ω_1 marginal	ω_2 backward	ω_3 regime-matched	Family	$\Delta^*_{1,3}$ [95% CI]
O1	$-.03233[-.04016, -.02450]$	$-.04116[-.04802, -.03430]$	$-.01912[-.02604, -.01221]$	R	$-.01320[-.01920, -.00683]$
O2	$-.03788[-.04368, -.03208]$	$-.03390[-.03931, -.02850]$	$-.01668[-.02213, -.01124]$	R	$-.02119[-.02692, -.01545]$
O3	$-.03566[-.04142, -.02990]$	$-.04352[-.04787, -.03917]$	$-.01639[-.02068, -.01210]$	R	$-.01927[-.02425, -.01396]$
U1	$-.02374[-.03025, -.01722]$	$-.03445[-.04045, -.02845]$	$-.00422[-.01073, +.\mathbf{00229}]$	U	$-.01952[-.02574, -.01391]$
U2	$-.04644[-.05215, -.04073]$	$-.03857[-.04344, -.03370]$	$-.00451[-.00957, +.\mathbf{00055}]$	U	$-.04193[-.04717, -.03661]$
U3	$-.02429[-.02844, -.02014]$	$-.02787[-.03150, -.02424]$	$-.00466[-.00852, -.00081]$	R	$-.01963[-.02358, -.01568]$
T1	$-.03431[-.04241, -.02620]$	$-.03330[-.04077, -.02584]$	$-.01929[-.02535, -.01323]$	R	$-.01501[-.02179, -.00811]$

A.3 Prespecified design and negative controls

Train/test sizes are O1 $860/901$, O2 $1761/804$, O3 $2573/689$, U1 $969/1007$, U2 $1976/1007$, U3 $2983/1007$, and T1 $3262/169$. The prespecified target menu contained 3-day direction, 3-day abnormal return versus QQQ, 3-day cross-sectional rank, and next-10-day realized-volatility regime; only volatility regime passed (mean joint AUC 0.582). The selected binary label is one when forward 10-day realized volatility strictly exceeds trailing 20-day volatility at the reference date; ties map to class 0.

For the text branch, articles from the prior 10 calendar days are concatenated, truncated to 4000 characters, encoded with `sentence-transformers/all-MiniLM-L6-v2` (384 dimensions), and replaced by `(no text)` when absent. Train-only PCA retains 50 components with `random_state=13`, followed by train-only standardization. The numeric branch contains 63 within-window standardized log-return coordinates plus eight indicators: 5-day and 20-day mean returns, full-window and 10-day return standard deviations, log-price OLS slope, RSI-14/100, custom-EMA MACD, and last-close z-score. This yields 71 numeric dimensions and 121 joint inputs. The fitted probe is `LogisticRegression(max_iter=3000, C=1.0, random_state=seed)` with the scikit-learn defaults used by the reported runs (L_2 penalty, `lbfgs`, tolerance 10^{-4}, intercept, no class weighting); primary model seed is 13.

The reported experiments used Python 3.11.5, NumPy 1.26.4, pandas 2.2.3, SciPy 1.15.2, scikit-learn 1.2.2, sentence-transformers 5.6.1, and PyTorch 2.5.1+cu121.

Exact scalar-second-moment normalization leaves both diagnostic ω_1–ω_3 contrasts nonzero. In the controlled $\sigma = 0$ comparison, marginal and exact-conditional confidence-loss gaps are both positive, so no sign reversal occurs. Decision-axis ratios are $0.578[0.537, 0.621]$ (U2 numeric), $0.975[0.946, 1.005]$ (U2 text), $0.792[0.759, 0.826]$ (O1 numeric), and $1.025[0.988, 1.061]$ (O1 text). None changes the primary panel labels.