RANKING IS NOT ENOUGH: SHARED-THRESHOLD ROBUSTNESS IN AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image detectors are often deployed on images that have been compressed, resized, blurred, transmitted, or re-digitized. These transformations can shift detector scores even when their within-condition ranking remains informative. As a result, strong conditionwise detection does not guarantee that one fixed orientation and threshold will work across all deployment conditions. We formalize this distinction in two stages. First, we study whether detector evidence itself survives processing. Our scalar response, q_1, compares the image change caused by one application of a fixed reconstruction operator with the additional change caused by a second application. We derive a response-survival decomposition that tracks how clean real-versus-generated separation is retained or lost through processing distortion and reconstruction-processing interaction. Second, we characterize when all conditions admit one common decision rule. Exact threshold intersections determine shared-threshold feasibility, while total variation gives lower bounds on any decision restricted to the response. Empirically, we evaluate q_1 across two real-image sources, six generator families, and 32 processing conditions using independent calibration and confirmation. We then apply the same protocol to 58 publicly available detectors. The results show both regimes: some scores support a useful shared operating point, while many retain strong conditionwise ranking but fail under one fixed rule. For q_1, the best shared threshold remains close to chance in its worst condition. These findings separate response survival from threshold compatibility and show why both are necessary for robust detector deployment.

1 INTRODUCTION

An AI-generated image detector rarely operates on pristine files. Images may be compressed by platforms, resized by browsers, blurred by editors, repeatedly transmitted, or re-digitized through screenshots and print-and-scan pipelines. These transformations can alter detector evidence while leaving the visible content largely unchanged. As a result, the same scene can cross a detector's decision boundary after processing, and repeated transformations can compound the shift. Early detectors exploit convolutional artifacts, frequency signatures, and local pixel dependencies (Wang et al., 2020; Frank et al., 2020; Tan et al., 2024a). Modern benchmarks broaden evaluation to diverse generators and degraded images (Zhu et al., 2023), while controlled studies show that generator diversity, content alignment, format, and model selection can affect transfer conclusions (Park & Owens, 2025; Guillaro et al., 2025; Gulrajani & Lopez-Paz, 2021). This raises a practical deployment question: *after image transformations, can one detector response support one fixed decision rule?*

A simple score-location example shows why. Suppose real-image scores initially lie well above generated-image scores, so a midpoint threshold separates the two classes. After resizing, both distributions may shift below that threshold while preserving their relative ordering. The resized condition can therefore retain strong ranking even though the original decision rule fails. Another transformation may even reverse the useful score direction, making one orientation favorable for some conditions and harmful for others. Condition-specific AUROC captures within-condition ranking but is insensitive to a common score shift. A shared deployment rule is stricter: it must use one orientation and one threshold across all conditions. Its success therefore depends not only on ranking, but also on absolute score locations and tails. Local ranking and shared-rule feasibility are thus different properties, and strong condition-specific structure need not imply a useful common operating point. Fig. 1(a) illustrates this mismatch between conditionwise ranking and a fixed deployment rule.

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.

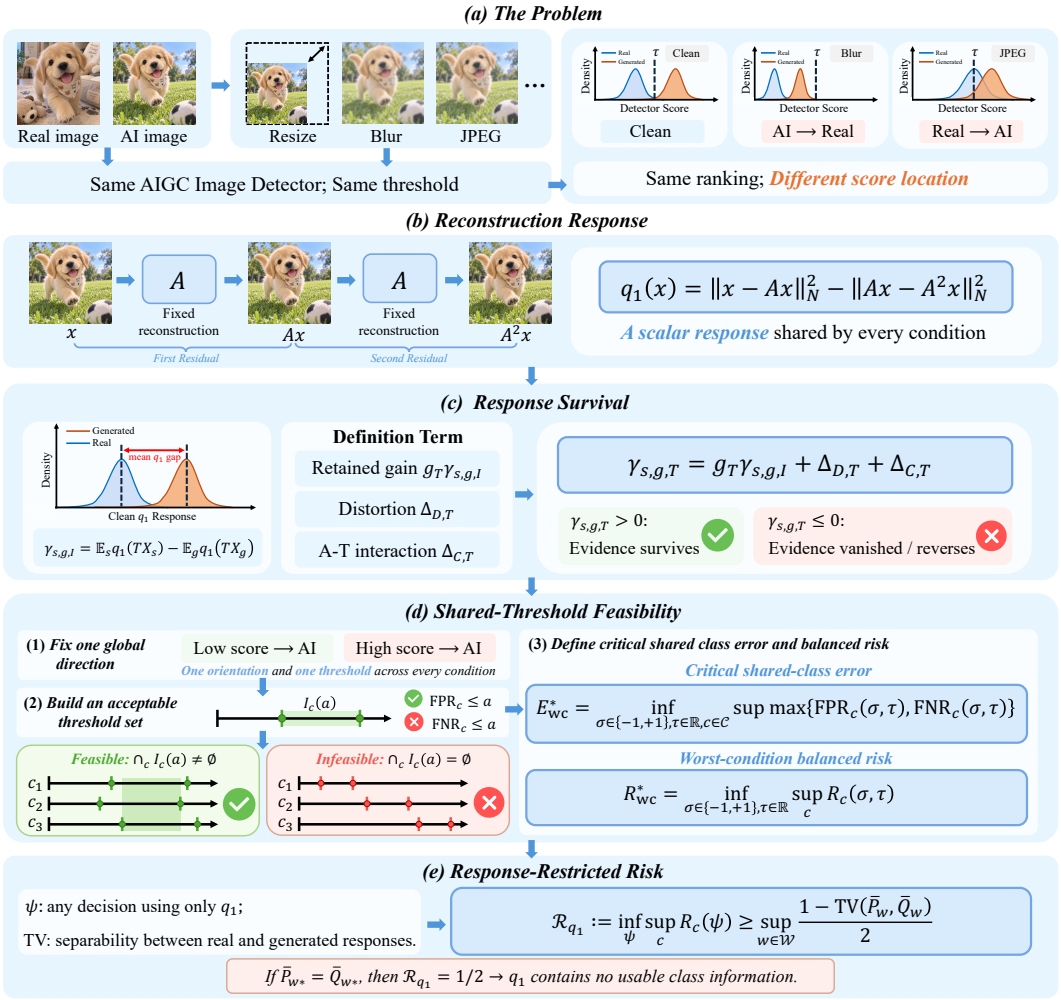


Figure 1: Response survival and shared-threshold feasibility: (a) Processing can shift score location or orientation. (b) q_1 compares successive reconstruction changes. (c) Processed separation decomposes into retained clean separation, distortion, and reconstruction–processing interaction. (d) A shared rule requires one orientation and threshold across all conditions. (e) Total variation lower-bounds response-restricted risk.

We summarize this conflict by the critical shared class error E^*_{wc}, defined by optimizing over both orientations and all thresholds. It is the smallest achievable value of the worst false-positive or false-negative rate across all conditions. Thus, $E^*_{\text{wc}} = 0.52$ means that every fixed rule misclassifies at least 52% of one class in some condition.

We study this question through a fixed reconstruction operator A. For an image x, the scalar response q_1 compares the squared change from x to its first reconstruction Ax with the additional change from Ax to a second reconstruction A^2x, as illustrated in Fig. 1(b). Intuitively, q_1 measures whether reconstruction largely settles after one application or continues to change at the next. Reconstruction and inversion already provide useful detection signals (Wang et al., 2023; Ricker et al., 2024; Cazenavette et al., 2024), while natural-image and spectral responses provide complementary evidence (Zhang et al., 2025c; Karageorgiou et al., 2025). Our first theoretical layer characterizes whether the real–generated response gap survives image processing, a property we call *response survival*. Fig. 1(c) summarizes this decomposition: clean first- and second-moment geometry provides the initial separation, while deterministic and stochastic processing modify it through retained gain, centered distortion, approximation error, and reconstruction–processing interaction. The interaction captures non-commutation: reconstructing a processed image need not match processing a

reconstructed image. Fig. 1(d–e) then connects the surviving response to shared-threshold feasibility and response-restricted risk. Appendix A expands this framework. Our contributions are:

- We formulate response survival and shared-threshold feasibility for AI-generated image detection and derive processing-aware separation bounds for fixed reconstruction responses.
- We characterize critical shared class error, shared-threshold feasibility, and response-restricted risk through threshold intersections, bounds, least-favorable mixtures, and condition coverage.
- We evaluate q_1 and 58 publicly available detector scores with independent calibration and confirmation across 384 source–generator–processing cells.

2 Related Work

AI-generated image detection and robustness. AI-generated image detectors use convolutional artifacts, frequency cues, local pixel statistics, reconstruction error, and frozen visual representations (Wang et al., 2020; Frank et al., 2020; Tan et al., 2024a; Wang et al., 2023; Ricker et al., 2024; Ojha et al., 2023). Recent benchmarks and detector designs emphasize transfer across generators, degraded images, and changing data distributions (Zhu et al., 2023; Park & Owens, 2025; Guillaro et al., 2025; Zhang et al., 2025a; Yang et al., 2025). These studies establish that detector ranking can generalize beyond a clean training distribution, but condition-specific AUROC does not determine whether one absolute score orientation and threshold remain valid across conditions. Our focus is this stronger deployment requirement: whether detector evidence survives processing and supports one shared operating rule.

Shared decisions under distribution shift. Reconstruction-based and fixed-response detectors motivate our use of a scalar response whose behavior can be followed through image transformations (Cazenavette et al., 2024; Zhong et al., 2025; Zhang et al., 2025c; Karageorgiou et al., 2025). Our analysis is also related to domain generalization and statistical decision theory (Ben-David et al., 2010; Blanchard et al., 2021; Garg et al., 2021; Neyman & Pearson, 1933; Huber & Strassen, 1973; Le Cam, 1986; Raginsky, 2016). Rather than studying average transfer or allowing condition-specific adaptation, we require one decision rule across the full condition set. We therefore separate three objects: whether the response itself survives processing, whether conditionwise response distributions admit a common threshold, and what risk remains for decisions restricted to that response.

3 Separation and Decision Theory

We organize the theory around the deployment sequence in Fig. 1: define a scalar response, track whether its class separation survives processing, determine whether the resulting distributions admit one shared threshold, and finally quantify the risk and coverage of decisions restricted to that response. Appendix B provides proofs and extensions, with notation summarized in Appendix B.1.

3.1 Problem Formulation

Let $\mathcal{X} = [0,1]^N$ denote canonical RGB images with normalized inner product $\langle u, v\rangle_N = N^{-1}u^\top v$ and norm $\|u\|_N^2 = \langle u, u\rangle_N$. A fixed reconstruction operator $A : \mathcal{X} \to \mathcal{X}$ induces the one-step reconstruction response

$$q_1(x) = \|x - Ax\|_N^2 - \|Ax - A^2x\|_N^2. \tag{1}$$

For real source s, generator g, and deterministic processing map T, define $\Delta_{s,g,T} = \mathbb{E}_g q_1(TX_g) - \mathbb{E}_s q_1(TX_s)$ and $\gamma_{s,g,T} = -\Delta_{s,g,T}$. Positive $\gamma_{s,g,T}$ means that the real mean response exceeds the generated mean response. For a stochastic channel K, the same definition uses a class-independent draw from $K(X_z, \cdot)$ for each class z. The response-survival framework links three questions: how clean response geometry produces separation (Appendix B.2), how processing changes that separation, and whether one shared fixed threshold rule controls the worst-condition risk. Strong data-processing inequalities provide complementary information-contraction results for common channels (Raginsky, 2016); our object tracks a specified scalar reconstruction response and its interaction with A.

3.2 Separation after Deterministic and Stochastic Processing

With the response fixed, we first characterize how deterministic and stochastic processing change the real–generated separation. Let $\rho(u, v) = \|u - v\|_N^2$. On a bounded domain $\mathcal{D}$ of diameter $D_{\mathcal{D}}$,

$|\rho(u,v)-\rho(u',v')| \le L_\rho(\|u-u'\|_N + \|v-v'\|_N)$ with $L_\rho = 2D_{\mathcal{D}}$. Assume $\mathcal{D}$ is invariant under A and the processing map, so all endpoints of the two reconstruction orbits remain inside it. For deterministic class-independent T, transport the clean orbit through T by $\widetilde{q}_T(x) = \rho(Tx, T(Ax)) - \rho(T(Ax), T(A^2x))$. Given $g_T \ge 0$, define $D_{T,g_T}(x) = \widetilde{q}_T(x) - g_T q_1(x)$. Let $\widehat{D}_{T,g_T}$ be a measurable approximation, and require one scalar b_T shared across classes such that $\max_{z\in\{s,g\}} \mathbb{E}_z|\widehat{D}_{T,g_T}(X_z) - b_T| \le \delta_T$. Define $\xi_T = \sum_z \mathbb{E}_z|D_{T,g_T}(X_z) - \widehat{D}_{T,g_T}(X_z)|$.

The reconstruction and processing maps interact through two levels. Put $k_T(x) = \|A(Tx) - T(Ax)\|_N$, $e_{2,T}(x) = \|A^2(Tx) - T(A^2x)\|_N$, and $\kappa_T = L_\rho \sum_{z\in\{s,g\}} \mathbb{E}_z[2k_T(X_z) + e_{2,T}(X_z)]$. If A is L_A-Lipschitz, then $e_{2,T}(x) \le L_A k_T(x) + k_T(Ax)$.

Let $c_T(x) = q_1(Tx) - \widetilde{q}_T(x)$. The response identity is

$$q_1(Tx) = g_T q_1(x) + D_{T,g_T}(x) + c_T(x). \tag{2}$$

Define the signed class contrasts

$$\Delta_{D,T} = \mathbb{E}_s D_{T,g_T}(X_s) - \mathbb{E}_g D_{T,g_T}(X_g), \qquad \Delta_{C,T} = \mathbb{E}_s c_T(X_s) - \mathbb{E}_g c_T(X_g). \tag{3}$$

Theorem 3.1 (Deterministic response decomposition and separation)**.** *Assume the displayed expectations are finite. Then*

$$\boxed{\gamma_{s,g,T} = g_T \gamma_{s,g,I} + \Delta_{D,T} + \Delta_{C,T}.} \tag{4}$$

If $\gamma_{s,g,I} \ge \gamma_{\mathrm{clean},L} > 0$, $g_T \ge g_L \ge 0$, $\Delta_{D,T} \ge d_L$, and $\Delta_{C,T} \ge c_L$ hold on one simultaneous event, then

$$\gamma_{s,g,T} \ge g_L \gamma_{\mathrm{clean},L} + d_L + c_L. \tag{5}$$

Under the preceding centered approximation and endpoint conditions, the choices $d_L = -2\delta_U - \xi_U$ and $c_L = -\kappa_U$ are valid and give

$$\boxed{\gamma_{s,g,T} \ge L_T := g_L \gamma_{\mathrm{clean},L} - 2\delta_U - \xi_U - \kappa_U.} \tag{6}$$

A positive right hand side certifies positive mean separation for the declared map and class pair. Simultaneous assumptions yield the corresponding uniform statement over the declared index set.

The proof, processing-specific bounds, and stochastic extension appear in Appendices B.3, B.4, and B.9, respectively.

3.3 One Shared Fixed Threshold Rule

Response survival is only a mean-level prerequisite: even a positive processed gap need not place all conditionwise distributions on a common threshold scale. Let $c = (s, g, T) \in \mathcal{C}$ and let $F_{g,c}$ and $F_{s,c}$ be the population distribution functions of $q_1(TX_g)$ and $q_1(TX_s)$. A rule output of one means generated. For $\sigma \in \{-1, +1\}$ and threshold τ, define $h_{\sigma,\tau}(z) = \mathbf{1}\{\sigma(z-\tau) < 0\}$. Equality at τ is assigned to the real class. For any class error level $a \in [0,1]$, define $I_c^+(a) = \{\tau : 1 - F_{g,c}(\tau^-) \le a,\ F_{s,c}(\tau^-) \le a\}$ and $I_c^-(a) = \{\tau : F_{g,c}(\tau) \le a,\ 1 - F_{s,c}(\tau) \le a\}$.

Theorem 3.2 (Composite detection by one threshold)**.** *A shared threshold rule whose two class errors are at most a in every declared cell exists exactly when*

$$\bigcap_c I_c^+(a) \ne \varnothing \qquad \textit{or} \qquad \bigcap_c I_c^-(a) \ne \varnothing. \tag{7}$$

The proof is given in Appendix B.5.

The first error level at which either intersection becomes nonempty is the critical shared class error

$$E_{\mathrm{wc}}^* = \inf_{\sigma\in\{-1,+1\},\tau\in\mathbb{R}} \sup_{c\in\mathcal{C}} \max\{\mathrm{FPR}_c(\sigma,\tau), \mathrm{FNR}_c(\sigma,\tau)\}. \tag{8}$$

This quantity traces feasibility without selecting a success cutoff: $E_{\mathrm{wc}}^* \le a$ exactly when the intersection in Theorem 3.2 is nonempty at level a. A value above one half means that every shared orientation and threshold misclassifies a majority of at least one class in at least one cell. A value below one half means that some shared rule keeps both class errors below one half in every cell.

$$R^*_{\mathrm{wc}} := \inf_{\sigma\in\{-1,+1\},\tau\in\mathbb{R}} \sup_c R_c(\sigma,\tau), \tag{9}$$

Let $\widehat{E}^*_{\mathrm{wc}}$ denote the empirical analogue obtained from the complete confirmatory threshold class. If every class conditional empirical distribution function lies within ε of its population counterpart, then

$$\max\{0, \widehat{E}^*_{\mathrm{wc}} - \varepsilon\} \le E^*_{\mathrm{wc}} \le \min\{1, \widehat{E}^*_{\mathrm{wc}} + \varepsilon\}. \tag{10}$$

The same simultaneous event covers every condition, orientation, and threshold, so the bound remains valid when methods are compared within one jointly controlled family (Appendix B.10).

3.4 Response-Restricted Lower Bounds and Condition Covers

A shared threshold is only one response-based decision class. We next characterize what can be achieved by any decision observing only q_1, and how guarantees extend from represented to intended conditions. Let P_c and Q_c denote the real and generated laws of the scalar response in cell c. A randomized response decision is a measurable $\psi : \mathbb{R} \to [0,1]$, where $\psi(z)$ is the probability of declaring generated. Under equal class priors, $R_c(\psi) = \frac{1}{2}\int \psi\, dP_c + \frac{1}{2}\int (1-\psi)\, dQ_c$. This is a statistical experiment observed only through q_1 in the sense of decision theory (Le Cam, 1986); least-favorable representatives connect it to composite robust testing (Huber & Strassen, 1973), while simple-hypothesis likelihood-ratio structure follows the Neyman–Pearson principle (Neyman & Pearson, 1933).

Theorem 3.3 (Bayes response risk and least-favorable mixtures)**.** *For every response decision ψ,*

$$R_c(\psi) \ge \frac{1 - \mathrm{TV}(P_c, Q_c)}{2}, \tag{11}$$

whereas the Bayes optimum satisfies

$$R^*_c := \inf_\psi R_c(\psi) = \frac{1 - \mathrm{TV}(P_c, Q_c)}{2}. \tag{12}$$

Let $\mathcal{W}$ be a declared set of probability vectors on the cell index and put $\overline{P}_w = \sum_c w_c P_c$ and $\overline{Q}_w = \sum_c w_c Q_c$. Then

$$\mathcal{R}_{q_1} := \inf_\psi \sup_c R_c(\psi) \ge \sup_{w\in\mathcal{W}} \frac{1 - \mathrm{TV}(\overline{P}_w, \overline{Q}_w)}{2}. \tag{13}$$

If an admissible w^* satisfies $\overline{P}_{w^*} = \overline{Q}_{w^*}$, then $\mathcal{R}_{q_1} = 1/2$.

Appendix B.6 proves these bounds and develops condition-cover guarantees.

3.5 Evaluation Objective

The primary objectives are the critical shared class error E^*_{wc} and population minimax balanced risk R^*_{wc} over one orientation and threshold shared across $\mathcal{C}$. Calibration selects a candidate operating point, while independent confirmation evaluates both that rule and the complete threshold class with simultaneous empirical-distribution bounds.

These objectives connect the preceding theory to evaluation. Processed separation tests whether response evidence survives processing, threshold intersections test whether conditionwise distributions admit a common rule, and response-restricted risk characterizes what remains achievable from the scalar response. Additional coverage and processing-boundary diagnostics are given in the Appendix.

4 Experimental Design

The experimental design mirrors the theoretical pipeline above. We estimate response survival, evaluate the complete shared-threshold class, and compare local and shared-rule behavior on independent confirmatory data. Appendix C provides the detailed protocol.

Evaluation scope. We use two real-image sources, six generator families, and 32 processing conditions, yielding 384 source–generator–processing cells. The real sources are AIGC natural images and AIGI StyleGAN3 images; generated images come from ADM, GLIDE, Midjourney, Stable Diffusion, StyleGAN3, and Wukong. Processing includes clean, Gaussian blur, blur–resize, and JPEG conditions (the complete registry is in Appendix D.3). This design follows cross-generator, degraded-image, repeated-transmission, and common-corruption evaluation settings (Zhu et al., 2023; Li et al., 2025a; Hendrycks & Dietterich, 2019).

Responses and controls. The primary response is q_1. Controlled alternatives include $|q_1|$, reconstruction error e_0, a frequency-phase response, ConvNeXt (Liu et al., 2022), DINOv2 (Oquab et al., 2023), DINOv3 (Siméoni et al., 2025), CLIP (Radford et al., 2021), and RealStats-style real-population statistics (Zisman & Shaham, 2026). All responses are computed from the same canonical images and condition assignments, allowing direct comparison across reconstruction, spectral, convolutional, self-supervised, and vision-language representations (Appendix C.1).

Shared-rule evaluation. For q_1, we evaluate both prespecified orientations, all 196,275 distinct threshold decisions per orientation, and the two exterior constant rules. Balanced risk averages false-positive and false-negative rates within each cell before taking the worst condition. Critical shared class error instead takes the larger class error within each cell and then minimizes the worst value over orientation and threshold. Condition-specific AUROC measures local ranking, while response atlases retain signed separation. Source, generator, and processing identities remain separate throughout inference so that incompatible score locations are not hidden by early aggregation.

Simultaneous inference. The primary q_1 analysis contains $m = 768$ class-conditional empirical distribution functions with $n = 768$ confirmatory observations each. At $\alpha = 0.0125$, the two-sided Dvoretzky–Kiefer–Wolfowitz union bound gives

$$\varepsilon_{\mathrm{DKW}} = \sqrt{\frac{\log(2m/\alpha)}{2n}} = 0.087347,$$

yielding simultaneous coverage of at least 0.9875 over the complete threshold class, E^*_{wc}, and R^*_{wc} (Massart, 1990). Controlled comparisons use a common max-t reference distribution, while exact and unavailable endpoints remain distinct (Appendix D.4).

Response-survival estimation. We evaluate the decomposition on four representative maps: clean, Gaussian blur condition 6, blur–resize condition 6, and JPEG quality 65. Calibration selects the nonnegative gain g and common distortion center b_T. Independent confirmatory ranks estimate the retained clean margin, signed distortion contrast, reconstruction–processing interaction, and processed margin using equal mixtures of the two real sources and six generators. Two thousand group-stratified bootstrap replicates provide descriptive 95% intervals. For this exact calculation, $\xi = 0$ and $\omega = 0$.

Calibration, confirmation, and public detectors. Ranks 1–256 within each source group form calibration and ranks 257–1,024 form confirmation. Calibration selects one orientation and threshold over all 384 cells; confirmation evaluates that rule together with the complete threshold class. We apply the same protocol to 58 publicly available detectors, covering 44,544 class-conditional distributions. Their simultaneous DKW radius is 0.101356 at familywise level 0.0125. Each paper contributes one prespecified method and checkpoint, fixed before the expanded response analysis; Appendix C.2 gives the inclusion and implementation details.

5 Results and Analysis

The results follow the same hierarchy. We first evaluate shared-rule feasibility and the operating point it selects, then test the response-survival decomposition, locate conditionwise incompatibility, and finally examine whether the same phenomenon persists across public detectors and controlled alternatives. Appendix D gives the detailed analyses.

Observation 1: Every shared q_1 threshold remains close to chance in its worst condition. We begin with the complete shared decision class. Table 1 evaluates both orientations, all 196,275 threshold decisions per orientation, and the two exterior constant rules. The simultaneous lower bound on minimax worst-condition balanced risk is 0.456326, limiting the advantage over chance to 4.37 percentage points. Fig. 2 displays the simultaneous lower and upper risk bounds at 17

Table 1: Worst condition risk over the complete shared threshold class.

Rule family	Orientations	Thresholds	Risk lower bound	Advantage ceiling	$E^\star$ [lower]
Shared fixed q_1 threshold	2	196,275	0.456326	0.043674	0.993490 [0.906142]

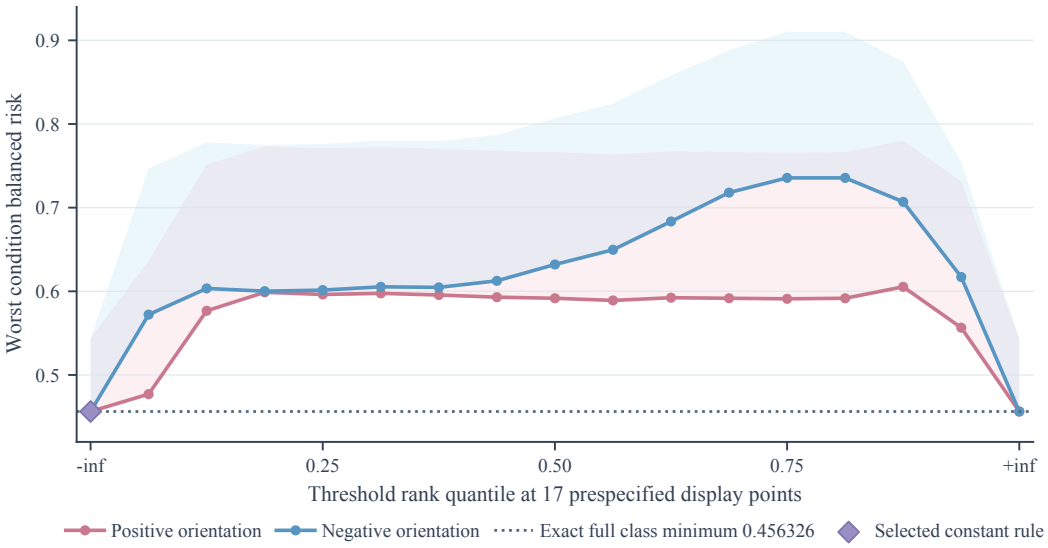


Figure 2: Representative threshold-rank samples of simultaneous worst-condition risk, together with the exact full-class minimum.

prespecified threshold-rank positions; its dotted line records the exact minimum over the complete class. The orientation curves separate across most interior ranks and meet at the constant decisions, illustrating why both orientations and endpoints must be included. The all-rule lower envelope characterizes attainable population performance, whereas the selected operating point describes the calibration outcome. Consequently, the exclusion does not depend on a chosen display grid or a favorable threshold sample: the finite display explains the risk geometry, while exhaustive enumeration and simultaneous uncertainty establish the bound (additional experimental summaries are in Appendix C.3).

Observation 2: Minimax calibration selects a constant rule with strongly asymmetric errors. We next examine the operating point selected without confirmatory responses. Calibration chooses the positive orientation at negative infinity, producing the constant all-real decision. Its simultaneous worst-condition balanced-risk interval is $[0.456326, 0.543674]$, with real-side false positives in $[0, 0.087347]$ and generated-side misses in $[0.912653, 1.0]$. Heterogeneous response locations therefore make a degenerate endpoint preferable to an interior shared threshold under the calibration objective. The complete classwise analysis reveals the same incompatibility: the critical shared class error is 0.993490 with simultaneous lower bound 0.906142. Its minimizing high-fake rule uses threshold 2.24737×10^{-4} and incurs false-positive rate 0.993490 on the AIGI StyleGAN3 real-image source under JPEG quality 35. Reversing the orientation increases the critical error to 0.996094, witnessed by Wukong generated images under JPEG qualities 35 and 50. Thus threshold movement and orientation reversal exchange which condition incurs the severe error, rather than resolving the conflict (Appendices D.1 and D.5).

Observation 3: Blur nearly erases the response margin, while JPEG preserves most of it. Turning to the response-survival layer, we test the decomposition in Fig. 3. The exact identity $\gamma_T = g\gamma_I + \Delta_D + \Delta_C$ from Eq. (4) closes to a maximum absolute discrepancy of 5.5×10^{-20}. Calibration selects g and b_T, while the independent confirmatory ranks estimate every contribution. Here Δ_D is the real-minus-generated mean contrast of $D_{T,g}$, and Δ_C is the contrast of $c_T = q_1 \circ T - \widetilde{q}_T$; the latter is the signed quantity bounded in absolute value by the distribution-free envelope κ_T in Theorem 3.1. The clean margin is 4.417×10^{-4}. Gaussian blur and blur–resize have gains 0.127 and

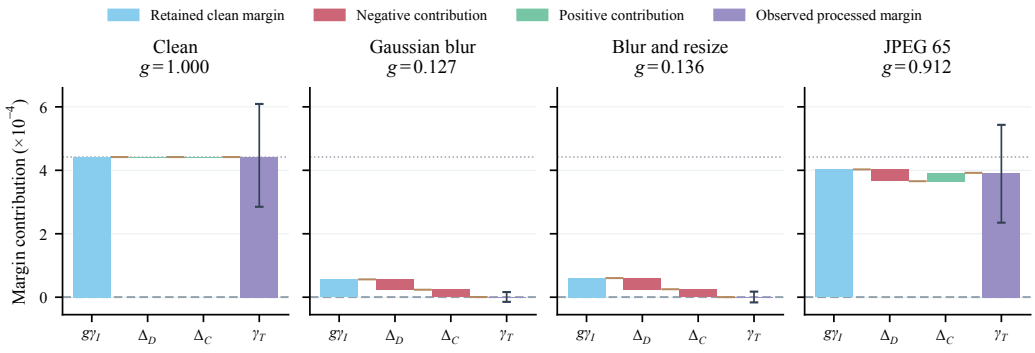


Figure 3: Response-survival decomposition across four processing conditions. Bars show retained margin, distortion, interaction, and processed margin; error bars give 95% bootstrap intervals.

0.136, retaining only 12.7% and 13.6% of that margin. Negative distortion and interaction contrasts consume 99.0% and 99.5% of the retained terms, leaving margins of 5.532×10^{-7} and 2.893×10^{-7} whose bootstrap intervals cross zero. By contrast, JPEG quality 65 has gain 0.912 and preserves a margin of 3.920×10^{-4}, or 88.7% of the clean margin, with interval $[2.349, 5.433] \times 10^{-4}$. The decomposition therefore distinguishes weak retention followed by cancellation from processing that preserves the original separation.

Observation 4: Useful conditionwise ranking does not imply one compatible threshold. We next connect mean-level response survival to distribution-level threshold compatibility. The response atlas in Fig. 4 locates the heterogeneity behind this distinction. Opposite signs across real-image sources or generators identify incompatible orientations even when processing changes little within a block. Condition-specific AUROC means range from 0.117886 to 0.695273; their lower endpoints range from 0.088462 to 0.649404, and their upper endpoints from 0.147313 to 0.741297. The upper part of this range confirms that q_1 can rank the classes effectively in individual conditions, while the low values can indicate a reversed local ordering. A global rule must nevertheless fix its orientation and threshold across every cell. Accordingly, the atlas identifies where directions and locations disagree, whereas the complete CDF analysis determines whether those distributions admit a shared operating point. Local ranking and shared-rule risk answer complementary questions and must be reported together; Appendices D.3 and D.6 provide the full response data and atlas interpretation.

Observation 5: Public detector scores exhibit the same threshold-compatibility bottleneck. Finally, Fig. 5 extends the comparison to 58 publicly available scores under the same population and independent splits. MIRROR is strongest by all three shared-rule summaries: its calibrated worst-condition risk is 0.256, its complete threshold class reaches point risk 0.245, and its minimum worst-class error is 0.483 with simultaneous lower bound 0.382. Across the collection, however, 53 simultaneous lower bounds on the critical class error exceed 0.5. These results show that strong conditionwise ordering can coexist with poor global decision compatibility, while also retaining the scores for which a shared operating point is useful. Because the images, condition membership, and splits are held fixed, the comparison attributes this variation to score compatibility within the evaluated protocol. Appendix D.2 summarizes the shared-rule results, and Appendix D.7 provides the complete 58-method table and further interpretation.

Observation 6: Controlled alternatives refute the registered direction criteria without resolving every mechanism. The controlled comparisons distinguish scientific outcomes from missing information. Absolute q_1, reconstruction error e_0, encoder-based natural statistics, and the phase response all have their prespecified all-condition direction criteria refuted. Exact-zero encoder and phase endpoints remain distinct from estimated small effects, while unavailable compatibility and coverage entries leave their mechanisms unresolved. Conversely, the supported repeated-transmission and re-digitization result concerns a paired degradation comparison under matched parents. It therefore complements the distribution-level shared-threshold result without establishing threshold compatibility. Table 5 reports the complete inventory, keeping unavailable endpoints distinct from evidence of robustness; Appendix D.4 details the decision criteria and endpoint diagnostics.

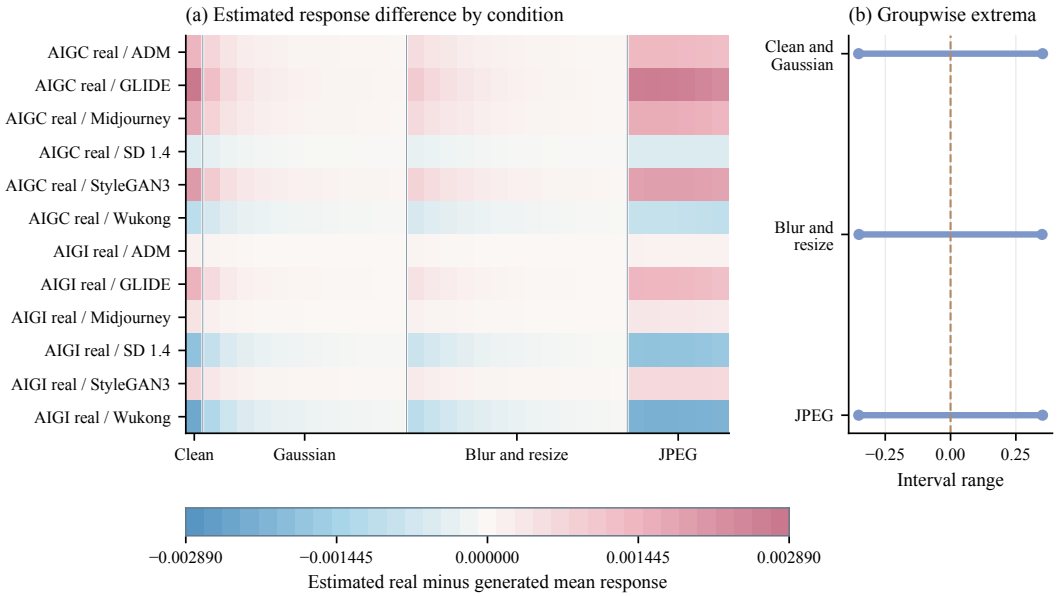


Figure 4: Estimated response differences and groupwise simultaneous interval extrema across clean-and-Gaussian, blur-resize, and JPEG conditions.

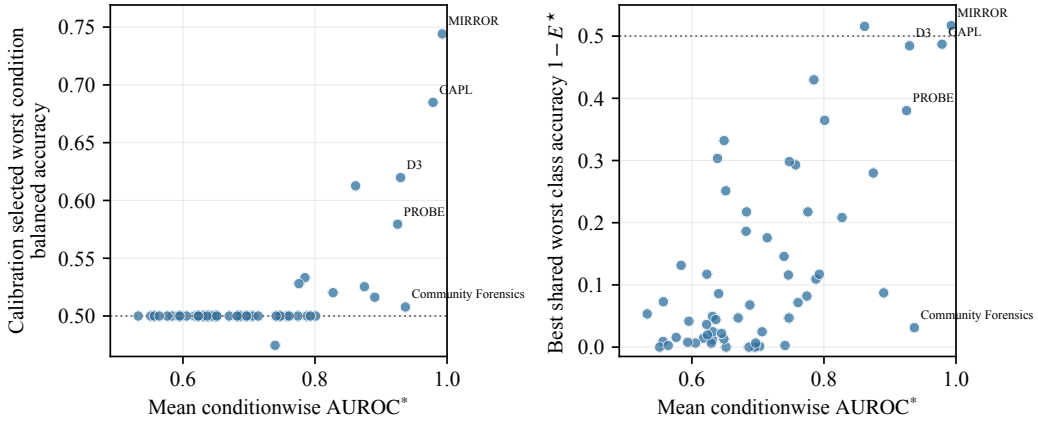


Figure 5: Conditionwise ranking versus shared-rule performance across 58 public detectors. Each horizontal coordinate averages $\mathrm{AUROC}^* = \max\{\mathrm{AUROC}, 1 - \mathrm{AUROC}\}$ over the 384 cells. Left: calibration-selected worst-condition balanced accuracy. Right: best shared worst-class accuracy over all orientations and thresholds.

6 Conclusion

We develop a decision-theoretic framework for AI-generated image detection that separates response survival from shared-threshold feasibility. A detector score may retain useful within-condition ranking after processing while still failing to support one portable operating point because score locations or useful orientations shift across conditions. Our response-survival decomposition identifies how clean separation is retained or lost under processing, while threshold intersections and response-restricted risk characterize the resulting decision limits. Across 384 source–generator–processing cells and 58 publicly available detectors, we observe both regimes: some scores admit a useful shared rule, whereas many preserve strong local ranking yet incur severe worst-condition errors under one fixed threshold. These results show that robust detector evaluation must distinguish whether evidence survives transformation from whether that surviving evidence remains compatible with one shared decision rule.

AI USAGE STATEMENT

Generative AI systems are used as limited editorial and implementation aids, including code drafting and debugging, language polishing, and manuscript organization. The authors design the mathematical framework and experiments, evaluate the scientific interpretations, and make the final decisions about every claim. The systems used to prepare the manuscript do not generate the experimental evidence or replace the analyses reported here. We inspect and test assisted code against the intended computations, check numerical statements against the recorded outputs, and review and revise all assisted prose. The authors therefore retain responsibility for the final text, mathematics, results, software, and accompanying materials.

REFERENCES

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024. doi: 10.1109/OJSP.2023.3337714.

Alexander Belton, Dominique Guillot, Apoorva Khare, and Mihai Putinar. Totally positive kernels, polya frequency functions, and their transforms. *Journal d'Analyse Mathématique*, 150(1):83–158, 2023. doi: 10.1007/s11854-022-0259-7.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175, 2010. doi: 10.1007/s10994-009-5152-4.

Roy Betser, Omer Hofman, Roman Vainshtein, and Guy Gilboa. General and domain-specific zero-shot detection of generated images via conditional likelihood. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.

Gilles Blanchard, Aniket Anand Deshmukh, Urun Dogan, Gyemin Lee, and Clayton Scott. Domain generalization by marginal transfer learning. *Journal of Machine Learning Research*, 22(2):1–55, 2021.

Zijie Cao, Weijie Tu, Yao Xiao, Weijian Deng, Liang Lin, and Pengxu Wei. Where detectors fail: Probing generative space for generalizable ai-generated image detection. In *Proceedings of the Forty-Third International Conference on Machine Learning*, 2026.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real world. *arXiv preprint arXiv:2406.09398*, 2024.

George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10759–10769, 2024.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7621–7639, 2024a.

Jiaxuan Chen, Jieteng Yao, and Li Niu. A single simple patch is all you need for ai-generated image detection. *arXiv preprint arXiv:2402.01123*, 2024b.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes ai-generated image detector easier generalizable. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Yingjian Chen, Lei Zhang, and Yakun Niu. ForgeLens: Data-efficient forgery focus for generalizable forgery image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16270–16280, 2025b.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. CO-SPY: Combining semantic and pixel features to detect synthetic images by AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13455–13465, 2025.

Sungik Choi, Hankook Lee, and Moontae Lee. Training-free detection of ai-generated images via cropping robustness. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. FIRE: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095167.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956. doi: 10.1214/aoms/1177728174.

Mingqi Fang, Ziguang Li, Lingyun Yu, Quanwei Yang, Hongtao Xie, and Yongdong Zhang. Forensic-MoE: Exploring comprehensive synthetic image detection traces with mixture of experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17772–17782, 2025.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3247–3258, 2020.

Vikas Garg, Adam Tauman Kalai, Katrina Ligett, and Steven Wu. Learn to expect the unexpected: Probably approximately correct domain generalization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3574–3582, 2021.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18685–18694, 2025.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Yuncheng Guo, Junyan Ye, Chenjue Zhang, Hengrui Kang, Haohuan Fu, Conghui He, and Weijia Li. OmniAID: Decoupling semantics and artifacts for universal ai-generated image detection in the wild. In *Proceedings of the Forty-Third International Conference on Machine Learning*, 2026.

Zhiyuan He, Pin-Yu Chen, and Tsung-Yi Ho. RIGID: A training-free and model-agnostic framework for robust ai-generated image detection. *arXiv preprint arXiv:2405.20112*, 2024.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Peter J. Huber and Volker Strassen. Minimax tests and the neyman-pearson lemma for capacities. *The Annals of Statistics*, 1(2):251–263, 1973. doi: 10.1214/aos/1176342363.

Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *IEEE International Conference on Image Processing*, pp. 3465–3469, 2022. doi: 10.1109/ICIP46576.2022.9897820.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution ai-generated image detection by spectral learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18706–18717, 2025.

Mamadou Keita, Wassim Hamidouche, Hessen Bougueffa Eutamene, Abdelmalik Taleb-Ahmed, and Abdenour Hadid. DeeCLIP: A robust and generalizable transformer-based framework for detecting ai-generated images. *arXiv preprint arXiv:2504.19876*, 2025.

Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. CLIPping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 1006–1015, 2024. doi: 10.1145/3652583.3658035.

Dahye Kim, Jaehyun Choi, Hyun Seok Seong, Seongho Kim, Donghun Lee, Sungwon Yi, and Jang-Ho Choi. Dissect and prune: Enhancing robustness in ai-generated image detection. In *Proceedings of the Forty-Third International Conference on Machine Learning*, 2026.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pp. 394–411, 2024. doi: 10.1007/978-3-031-73220-1_23.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer, New York, 1986. doi: 10.1007/978-1-4612-4946-7.

Chunxiao Li, Xiaoxiao Wang, Meiling Li, Boming Miao, Peng Sun, Yunjian Zhang, Xiangyang Ji, and Yao Zhu. Bridging the gap between ideal and real-world evaluation: Benchmarking ai-generated image detection in challenging scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20379–20389, 2025a.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2405–2414, 2025b.

Yiheng Li, Zichang Tan, Guoqing Xu, Zhen Lei, Xu Zhou, and Yang Yang. Towards generalizable ai-generated image detection via image-adaptive prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Shuqiao Liang, Jian Liu, Renzhang Chen, and Quanlong Guan. FerretNet: Efficient synthetic image detection via local pixel dependencies. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ruiqi Liu, Manni Cui, Ziheng Qin, Zhiyuan Yan, Ruoxin Chen, Yi Han, Zhiheng Li, Junkai Chen, ZhiJin Chen, Kaiqing Lin, Jialiang Shen, Lubin Weng, Jing Dong, Yan Wang, and Shu Wu. MIRROR: Manifold ideal reference reconstructor for generalizable ai-generated image detection. *arXiv preprint arXiv:2602.02222*, 2026.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Sara Mandelli, Paolo Bestagini, and Stefano Tubaro. When synthetic traces hide real content: Analysis of stable diffusion image laundering. *arXiv preprint arXiv:2407.10736*, 2024.

Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990. doi: 10.1214/aop/1176990746.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231(694–706):289–337, 1933. doi: 10.1098/rsta.1933.0009.

Tai D. Nguyen, Aref Azizpour, and Matthew C. Stamm. Forensic self-descriptions are all you need for zero-shot detection, open-set source attribution, and clustering of ai-generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3040–3050, 2025.

Jun Nie, Yonggang Zhang, Tongliang Liu, Yiu ming Cheung, Bo Han, and Xinmei Tian. Epistemic uncertainty for generated image detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8245–8257, 2025.

Haotian Qin, Dongliang Chang, Yueying Gao, Bingyao Yu, Lei Chen, and Zhanyu Ma. Multimodal conditional information bottleneck for generalizable ai-generated image detection. *arXiv preprint arXiv:2505.15217*, 2025.

Ziheng Qin, Yuheng Ji, Renshuai Tao, Yuxuan Tian, Yuyang Liu, Yipu Wang, and Xiaolong Zheng. Scaling up ai-generated image detection with generator-aware prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Maxim Raginsky. Strong data processing inequalities and phi-sobolev inequalities for discrete channels. *IEEE Transactions on Information Theory*, 62(6):3355–3389, 2016. doi: 10.1109/TIT.2016.2549542.

Anirudh Sundara Rajan and Yong Jae Lee. Stay-positive: A case for ignoring real image features in fake image detection. In *Proceedings of the Forty-Second International Conference on Machine Learning*, 2025.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *International Conference on Learning Representations*, 2025.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9130–9140, 2024.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023. doi: 10.1145/3576915.3616588.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12105–12114, 2023.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024a.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space learning. *arXiv preprint arXiv:2403.07240*, 2024b.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2P-CLIP: Injecting category common prompt in CLIP to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025.

Nilesh Tripuraneni, Michael I. Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

M. Voorhoeve. A generalization of descartes' rule. *Journal of the London Mathematical Society*, s2-20(3):446–456, 1979. doi: 10.1112/jlms/s2-20.3.446.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8695–8704, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023.

Haiwei Wu, Jiantao Zhou, and Shile Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *IEEE Transactions on Artificial Intelligence*, pp. 1–11, 2025. doi: 10.1109/TAI.2025.3641104.

Jiazhen Yan, Ziqiang Li, Fan Wang, Boyu Wang, Ziwen He, and Zhangjie Fu. DGS-Net: Distillation-guided gradient surgery for CLIP fine-tuning in ai-generated image detection. *arXiv preprint arXiv:2511.13108*, 2025a.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, 2025b.

Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. D3: Scaling up deepfake detection by learning from discrepancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23850–23859, 2025.

Lin Yuan, Xiaowan Li, Yan Zhang, Jiawei Zhang, Hongbo Li, and Xinbo Gao. MLEP: Multi-granularity local entropy patterns for universal ai-generated image detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Boxuan Zhang, Jianing Zhu, Qifan Wang, Jiang Liu, and Ruixiang Tang. Micro-defects expose macro-fakes: Detecting ai-generated images via local distributional shifts. *arXiv preprint arXiv:2605.09296*, 2026.

Haifeng Zhang, Qinghui He, Xiuli Bi, Weisheng Li, Bo Liu, and Bin Xiao. Towards universal ai-generated image detection by variational information bottleneck network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23828–23837, 2025a.

Yanran Zhang, Bingyao Yu, Yu Zheng, Wenzhao Zheng, Yueqi Duan, Lei Chen, Jie Zhou, and Jiwen Lu. D3QE: Learning discrete distribution discrepancy-aware quantization error for autoregressive-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16292–16301, 2025b.

Yonggang Zhang, Jun Nie, Xinmei Tian, Mingming Gong, Kun Zhang, and Bo Han. Detecting generated images by fitting natural image distributions. In *Advances in Neural Information Processing Systems*, volume 38, 2025c. doi: 10.52202/085713-1052.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. PatchCraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.

Nan Zhong, Haoyu Chen, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Beyond generation: A diffusion-based low-level feature extractor for detecting ai-generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8258–8268, 2025.

Nan Zhong, Mian Zou, Yiran Xu, Zhenxing Qian, Xinpeng Zhang, Baoyuan Wu, and Kede Ma. Self-supervised ai-generated image detection: A camera metadata perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Xiaoyu Zhou, Jianwei Fei, Peipeng Yu, Jingchang Xie, Chong Cheng, and Zhihua Xia. PGC: Peak-guided calibration for generalizable ai-generated image detection. In *Proceedings of the Forty-Third International Conference on Machine Learning*, 2026.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, 2023. doi: 10.52202/075280-3398.

Haim Zisman and Uri Shaham. Realstats: A rigorous real-only statistical framework for fake image detection. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 892–900, 2026.

A Response Survival and Shared-Threshold Feasibility

The second layer turns response distributions into decisions. For each source, generator, and processing condition, every candidate class error level defines an atom-aware set of acceptable thresholds for each orientation. *Shared-threshold feasibility* holds exactly when these sets intersect within one orientation. The first feasible level is the critical shared class error. Simultaneous empirical-distribution bands bound this quantity and every distinct threshold decision, including the two constant endpoints (Dvoretzky et al., 1956; Massart, 1990). Beyond thresholds, total variation yields the Bayes *response-restricted risk* for decisions observing only q_1, and least-favorable mixtures lower-bound the worst condition. A condition-cover result connects represented conditions through a decision-class risk pseudometric, while a missing-response-cluster construction captures the chance-level limit when distinct unseen possibilities induce indistinguishable response laws. Finally, ordered spectral attenuation and perturbation control locate a *processing boundary*, the point at which response direction changes along an ordered processing path. These results connect information contraction, distribution shift, one observable scalar, and one shared decision.

Together, these layers provide a diagnostic sequence. A small clean margin points to response construction; large transport or non-commutation terms point to processing sensitivity; and separated conditionwise distributions with an empty threshold intersection point to incompatible locations or tails. Each outcome suggests a different response redesign, so the theory offers more guidance than one aggregate robustness score.

We connect this decomposition to observed responses on four representative conditions. Fig. 3 shows that Gaussian blur and blur and resize retain only 12.7% and 13.6% of the clean margin, after which distortion and reconstruction processing interaction consume nearly all of the retained contrast. JPEG quality 65 retains 88.7% of the clean margin. The same identity therefore explains both margin collapse and margin survival within one fixed response.

Our evaluation instantiates this question with two real-image sources, six generator families, and 32 image-processing conditions, forming 384 source, generator, and processing cells. Independent calibration and confirmatory responses separate operating-point selection from evaluation. We test both orientations and all 196,275 distinct q_1 threshold decisions per orientation. Table 1 summarizes the full threshold result. The simultaneous lower bound on minimax worst-condition balanced risk

Table 2: Notation for response survival and shared-rule risk.

Symbol	Meaning
$x \in \mathcal{X}$	Canonical RGB image in the bounded image domain.
A	Fixed reconstruction operator; $A^2 = A \circ A$.
T, K	Deterministic processing map and class-independent stochastic channel.
q_1	Difference between the first and second squared reconstruction-edge lengths.
$\gamma_{s,g,T}$	Real-minus-generated mean response separation after processing.
$\mathcal{C}$	Index set of evaluated source, generator, and processing conditions.
$h_{\sigma,\tau}$	Shared threshold rule with orientation σ and threshold τ.
R_c	Balanced risk in condition c.
$E^*_{\mathrm{wc}}, R^*_{\mathrm{wc}}$	Minimum worst class error and minimum worst balanced risk over all shared threshold rules.
$d_{\mathcal{H}}$	Risk pseudometric for a fixed response-decision class $\mathcal{H}$.
$\rho(J)$	Cover radius of a represented condition subset J.

is 0.456326, so every shared threshold has an advantage over chance of at most 4.37 percentage points. The point critical shared class error is 0.993490, with simultaneous lower confidence bound 0.906142. Thus every shared q_1 rule has at least one cell in which at least one class is misclassified at a population rate above 90.61% on the simultaneous event. At the same time, condition-specific AUROC reaches 0.695273 and the response atlases contain substantial signed structure. We then apply the identical rank split and shared-threshold decision class to 58 publicly available AI-generated image detection methods, producing 15,204,352 scores. Among 58 publicly available methods, 46 have point minimax balanced risk exactly 0.5. The median method reaches mean conditionwise orientation-free AUROC 0.685, while the median minimum worst class error over every shared orientation and threshold is 0.930. Simultaneous lower bounds exceed 0.5 for 53 methods. The resulting comparison separates a property of the proposed q_1 response from a broader property of released detector score scales. The 768 simultaneously controlled q_1 response distributions show why the distinction matters: favorable local ordering appears in some cells, yet their admissible cutoffs fail to overlap across the full set. The central failure is incompatibility of response directions and locations across conditions.

B Additional Theory and Proofs

The mathematical results have distinct premise layers. Exact response decompositions, shared-threshold intersections, and response-restricted Bayes identities need no local geometric approximation. Separation after processing requires a clean lower bound and simultaneous upper bounds on retained gain, centered distortion, approximation error, and non-commutation. Local Taylor geometry, codec endpoint bounds, alias-free subspaces, and ordered spectral structure supply optional sufficient conditions for those terms. The separation, processing, cover, and spectral results are conditional mathematical statements; empirical conclusions require source-bound estimates of their premises.

B.1 Notation

Table 2 collects the objects used throughout the theory before their first detailed construction.

B.2 Clean Separation from Local Geometry

Write $F_j(x) = A^j x - A^{j+1}x$ for $j \in \{0,1\}$. At a deterministic reference $r \in \mathcal{X}$, selected independently of class responses, define $g_j = F_j(r)$, $B_j = DF_j(r)$, $a = \|g_0\|_N^2 - \|g_1\|_N^2$, $\ell = 2(B_0^\top W g_0 - B_1^\top W g_1)$, and $Q = B_0^\top W B_0 - B_1^\top W B_1$, where $W = N^{-1} I_N$. The symmetric matrix Q measures the local quadratic contrast between the two reconstruction edges.

Lemma B.1 (Local quadratic enclosure)**.** *Let $r_{\mathrm{loc}} > 0$. Suppose that for every y in the Euclidean ball $\|y - r\|_2 \le r_{\mathrm{loc}}$, each F_j has a C^2 extension to an open set containing that closed ball and*

$$\|D^2 F_j(y)[u,v]\|_N \le L_j \|u\|_2 \|v\|_2, \qquad j \in \{0,1\}, \tag{14}$$

and define $b_j = \|B_j\|_{2\to N}$ and $G_j = \|g_j\|_N$. Every h with $\|h\|_2 \le t \le r_{\mathrm{loc}}$ then satisfies

$$\left| q_1(r+h) - a - \ell^\top h - h^\top Q h \right| \le \Psi(r,t), \tag{15}$$

where

$$\Psi(r,t) = \sum_{j\in\{0,1\}} \left[L_j(G_j + b_j t)t^2 + \frac{L_j^2 t^4}{4} \right]. \tag{16}$$

Proof. Taylor's theorem gives $F_j(r+h) = g_j + B_j h + u_j(h)$ with $\|u_j(h)\|_N \le L_j\|h\|_2^2/2$. Hence

$$\left| \|F_j(r+h)\|_N^2 - \|g_j + B_j h\|_N^2 \right| \le 2\|g_j + B_j h\|_N \|u_j(h)\|_N + \|u_j(h)\|_N^2 \tag{17}$$

$$\le L_j(G_j + b_j t)t^2 + \frac{L_j^2 t^4}{4}. \tag{18}$$

Adding the absolute errors of the two orbit edges proves the claim. □

For $z \in \{s, g\}$, suppose $X_z = r + V_z + c_z$ almost surely, with $\|V_z\|_2 \le t_z$, $\|c_z\|_2 \le c_z^{\max}$, and $t_z + c_z^{\max} \le r_{\mathrm{loc}}$. Here c_z is a measurable sample-dependent deviation from the selected local coordinates. Define $C_z^{\max} = c_z^{\max}(\|\ell\|_2 + 2\|Q\|_{\mathrm{op}} t_z + \|Q\|_{\mathrm{op}} c_z^{\max})$, $\mu_z = \mathbb{E}[V_z]$, and $M_z = \mathbb{E}[V_z V_z^\top]$.

Theorem B.2 (Clean separation)**.** *Assume Lemma B.1 and the displayed local-support conditions. Define*

$$\Lambda_{s,g} = \ell^\top(\mu_g - \mu_s) + \mathrm{tr}(Q(M_g - M_s)) \tag{19}$$

and

$$\varepsilon_{s,g} = \sum_{z\in\{s,g\}} \left[\Psi(r, t_z + c_z^{\max}) + C_z^{\max} \right]. \tag{20}$$

If simultaneous one-sided bounds imply

$$\Lambda_{s,g} + \varepsilon_{s,g} \le -\gamma_{\mathrm{clean},L} < 0, \tag{21}$$

then

$$\mathbb{E}_s q_1(X_s) - \mathbb{E}_g q_1(X_g) \ge \gamma_{\mathrm{clean},L}. \tag{22}$$

If, on one simultaneous event, all support, derivative, moment, and error bounds hold and the premise holds for every $(s,g) \in \mathcal{S} \times \mathcal{G}$ with one common positive $\gamma_{\mathrm{clean},L}$, then the same lower bound is uniform over the declared index set.

Proof. Apply Lemma B.1 to $h = V_z + c_z$. Replacing the linear–quadratic expression at $V_z + c_z$ by its value at V_z costs at most $C_z^{\max}$. The constant a cancels between classes, and $\mathbb{E}[V_z^\top Q V_z] = \mathrm{tr}(QM_z)$. The generated-minus-real difference is at most $\Lambda_{s,g} + \varepsilon_{s,g}$, which proves Eq. (22). □

Two block conditions make the clean theorem interpretable. Let P_N be an orthogonal projection selected independently of class responses, put $P_C = I - P_N$, and write $Q_{UV} = P_U Q P_V$ and $\Delta M_{UV} = P_U(M_g - M_s)P_V$.

Corollary B.3 (Negative block with generated second-moment excess)**.** *Under the support and differentiability assumptions of Theorem B.2, suppose*

$$Q_{NN} \preceq -\alpha P_N, \qquad \Delta M_{NN} \succeq 0, \qquad \mathrm{tr}(\Delta M_{NN}) \ge \eta, \tag{23}$$

$$\|Q_{CC}\|_{\mathrm{op}} \le \beta, \qquad \|\Delta M_{CC}\|_* \le \zeta_C, \qquad \|Q_{NC}\|_{\mathrm{op}} \le \chi, \quad \|\Delta M_{NC}\|_* \le \zeta_X, \tag{24}$$

with $\alpha, \eta > 0$, and suppose $|\ell^\top(\mu_g - \mu_s)| \le b_\mu$ and $\varepsilon_{s,g} \le \varepsilon$. Then

$$\mathbb{E}_s q_1(X_s) - \mathbb{E}_g q_1(X_g) \ge \gamma_A := \alpha\eta - \beta\zeta_C - 2\chi\zeta_X - b_\mu - \varepsilon. \tag{25}$$

In particular, $\gamma_A > 0$ is a sufficient clean-separation condition.

Proof. The main block obeys $\mathrm{tr}(Q_{NN}\Delta M_{NN}) \le -\alpha\,\mathrm{tr}(\Delta M_{NN})$. Trace duality bounds the complement by $\beta\zeta_C$. Symmetry of Q and ΔM makes the two cross terms equal, and their total absolute value is at most $2\chi\zeta_X$. Add the linear and approximation costs and change from generated-minus-real to real-minus-generated sign. □

Let P_E be another response-independent orthogonal projection, put $P_\perp = I - P_E$, and define $H_E = P_E(M_s - M_g)P_E$.

Corollary B.4 (Positive block with real second-moment excess)**.** *Under the support and differentiability assumptions of Theorem B.2, suppose*

$$Q_{EE} \succeq \alpha_+ P_E, \qquad H_E \succeq 0, \qquad \operatorname{tr}(H_E) \geq \eta_+, \tag{26}$$

$$\|Q_{\perp\perp}\|_{\mathrm{op}} \leq \beta_+, \qquad \|\Delta M_{\perp\perp}\|_* \leq \zeta_+, \qquad \|Q_{E\perp}\|_{\mathrm{op}} \leq \chi_+, \quad \|\Delta M_{E\perp}\|_* \leq \xi_+, \tag{27}$$

and $|\ell^\top(\mu_g - \mu_s)| \leq b_+$ *and* $\varepsilon_{s,g} \leq \varepsilon_+$*. Then*

$$\mathbb{E}_s q_1(X_s) - \mathbb{E}_g q_1(X_g) \geq \gamma_B := \alpha_+\eta_+ - \beta_+\zeta_+ - 2\chi_+\xi_+ - b_+ - \varepsilon_+. \tag{28}$$

Thus $\gamma_B > 0$ *is a sufficient clean-separation condition.*

Proof. Since $\Delta M_{EE} = -H_E$, $\operatorname{tr}(Q_{EE}\Delta M_{EE}) \leq -\alpha_+ \operatorname{tr}(H_E)$. The complement and two cross terms cost at most $\beta_+\zeta_+ + 2\chi_+\xi_+$. Adding the linear and approximation costs gives the displayed real-minus-generated margin. □

The block conditions expose two mechanisms: generated images can place more second-moment mass in a negative response-active subspace, or real images can place more mass in a positive one. The obvious alternative is to infer the sign from a global reconstruction-error average. The block decomposition is chosen because it identifies the spectral region producing the sign and accounts explicitly for complement, cross-block, linear, and curvature costs.

The local enclosure also identifies how the assumptions can be assessed. The derivative envelopes L_j control the nonlinear remainder over a declared neighborhood, while t_z and $c_z^{\max}$ control how much class probability enters that neighborhood and how far the chosen coordinates depart from the image set. The leading contrast depends only on classwise first and second moments of V_z. Consequently, its constituents admit separate one-sided bounds: operator quantities can be bounded uniformly over the declared support, and moment quantities can be estimated from response-independent samples. Hoeffding-type concentration applies to bounded moment functionals (Hoeffding, 1963). The clean conclusion uses their joint event, so a positive margin means that the signed block contribution exceeds every displayed linear, complement, cross-block, and nonlinear cost.

A random reference is admissible when it has the same law in both classes. Conditioning on the reference produces the same enclosure with measurable integrable constants, and averaging cancels the reference-dependent constant term. This form can represent a reference drawn from an external real-image population. A class-dependent reference would change the leading contrast by an uncancelled constant, so the common-reference construction is part of the estimand. Likewise, the response-active projections are selected independently of the class responses used to estimate their moments. These choices separate geometric specification from empirical confirmation and keep the clean margin interpretable.

B.3 Additional Proofs for Response Survival

Restatement (Theorem 3.1: Deterministic response decomposition and separation)**.** Use the notation and the centered approximation, endpoint, and non-commutation conditions defined in the main text. Assume the displayed expectations are finite. Then

$$\boxed{\gamma_{s,g,T} = g_T\gamma_{s,g,I} + \Delta_{D,T} + \Delta_{C,T}.} \tag{4}$$

If $\gamma_{s,g,I} \geq \gamma_{\mathrm{clean},L} > 0$, $g_T \geq g_L \geq 0$, $\Delta_{D,T} \geq d_L$, and $\Delta_{C,T} \geq c_L$ hold on one simultaneous event, then

$$\gamma_{s,g,T} \geq g_L\gamma_{\mathrm{clean},L} + d_L + c_L. \tag{5}$$

Under the centered approximation and endpoint conditions, the choices $d_L = -2\delta_U - \xi_U$ and $c_L = -\kappa_U$ are valid and give

$$\boxed{\gamma_{s,g,T} \geq L_T := g_L\gamma_{\mathrm{clean},L} - 2\delta_U - \xi_U - \kappa_U.} \tag{6}$$

A positive right hand side certifies positive mean separation for the declared map and class pair. Simultaneous assumptions yield the corresponding uniform statement over the declared index set.

Proof of Theorem 3.1. Taking real minus generated expectations in Eq. (2) gives Eq. (4). The signed one-sided enclosures give Eq. (5). For the endpoint specialization, the common center cancels between class expectations and gives $\Delta_{D,T} \geq -2\delta_T - \xi_T$. Replacing $T(Ax)$ by $A(Tx)$ changes the first squared distance by at most $L_\rho k_T(x)$. Replacing both endpoints of the second squared distance costs at most $L_\rho[k_T(x) + e_{2,T}(x)]$. Summing over the two classes gives $\Delta_{C,T} \geq -\kappa_T$. Substitution proves Eq. (6). □

The simultaneous version takes a common clean lower bound together with conditionwise lower bounds on the signed contrasts, or conditionwise upper bounds on the endpoint envelopes, before minimizing the resulting expression over the desired index set. This order keeps every processing map visible. The framework permits source-specific couplings for a stochastic channel because the input laws differ, while the kernel itself remains class independent. Reporting the representative map and the coupling with L_K makes the transport interpretation reproducible.

B.4 Processing Specializations

For ideal block-DCT quantization, let $U \in \mathbb{R}^{N\times N}$ be orthogonal. For coefficient set $\mathcal{B}$ and steps $\Delta_j > 0$, define $T_Q x = \Pi_{\mathcal{X}}[U^\top Q_\Delta(Ux)]$ with $[Q_\Delta(v)]_j = \Delta_j \operatorname{round}(v_j/\Delta_j)$ for $j \in \mathcal{B}$. Orthogonality and projection yield $\|T_Q u - u\|_N \leq \epsilon_{Q,N} := (2\sqrt{N})^{-1}(\sum_{j\in\mathcal{B}} \Delta_j^2)^{1/2}$. Suppose both clean orbit edges satisfy $0 < r_{\min,L} \leq \|F_j(x)\|_N \leq r_{\max,U}$, put $h_{Q,U} = 2\epsilon_{Q,N}$, $g_{Q,L} = (1 - h_{Q,U}/r_{\min,L})^2$, and require $r_{\min,L} > h_{Q,U}$. Define $E_{Q,2,U} = 2h_{Q,U}r_{\max,U} + h_{Q,U}^2$ and $\overline{E}_{Q,U} = E_{Q,2,U} + (1 - g_{Q,L})r_{\max,U}^2$.

Corollary B.5 (Ideal DCT separation)**.** *Assume the displayed edge support, the bounded-domain conditions of Theorem 3.1, the simultaneous premise $\gamma_{s,g,I} \geq \gamma_{\mathrm{clean},L} > 0$, and a simultaneous upper bound κ_Q built from $k_Q(X_z)$ and $e_{2,Q}(X_z)$ for both classes. Then*

$$\gamma_{s,g,Q} \geq g_{Q,L}\gamma_{\mathrm{clean},L} - 4\overline{E}_{Q,U} - \kappa_Q. \tag{29}$$

Proof. The quantization errors at the two endpoints of an edge differ by at most $h_{Q,U}$. Reverse triangle inequality gives the squared gain, while expansion of the squared norm gives $E_{Q,2,U}$. Adding the difference between gains one and $g_{Q,L}$ yields $|\rho(T_Q u, T_Q v) - g_{Q,L}\rho(u,v)| \leq \overline{E}_{Q,U}$. The two orbit edges give $|D_{Q,g_{Q,L}}(x)| \leq 2\overline{E}_{Q,U}$. In Theorem 3.1, choose the exact distortion, $b_Q = 0$, $\delta_Q = 2\overline{E}_{Q,U}$, and $\xi_Q = 0$. □

For an exact practical JPEG map T_{jpg}, the declared recipe includes color conversion, chroma subsampling, block alignment, quality tables, level shifts, coefficient rounding, inverse-transform rounding, clipping, integer conversion, and RGB decoding. Suppose every orbit endpoint satisfies $\|T_{\mathrm{jpg}}u - u\|_N \leq \epsilon_{\mathrm{jpg},U}$. A telescoping comparison may bound this value by a sum of final-RGB errors for ideal quantization, color, subsampling, grid, inverse transform, clipping, and integer conversion. Set $h_{\mathrm{jpg},U} = 2\epsilon_{\mathrm{jpg},U}$, $g_{\mathrm{jpg},L} = (1 - h_{\mathrm{jpg},U}/r_{\min,L})^2$, and $\overline{E}_{\mathrm{jpg},U} = 2h_{\mathrm{jpg},U}r_{\max,U} + h_{\mathrm{jpg},U}^2 + (1 - g_{\mathrm{jpg},L})r_{\max,U}^2$.

For the telescoping comparison, define hybrid maps from the same canonical RGB input to the final canonical RGB output. Consecutive hybrids differ in exactly one codec component, with all later components already applied. Each summand bounds the supremum of the final-RGB difference between adjacent hybrids. An intermediate-norm estimate is multiplied by the explicit Lipschitz gains of all later components. This construction accommodates the discontinuities introduced by rounding and clipping through cellwise endpoint bounds.

Corollary B.6 (Practical JPEG separation)**.** *Assume the bounded invariant-domain and simultaneous clean-separation premises of Theorem 3.1, with $\gamma_{s,g,I} \geq \gamma_{\mathrm{clean},L} > 0$, for the exact practical map. If $r_{\min,L} > h_{\mathrm{jpg},U}$ and this map has simultaneous two-level non-commutation bound*

$$\kappa_{\mathrm{jpg},U} \geq L_\rho \sum_{z\in\{s,g\}} \mathbb{E}_z[2k_{\mathrm{jpg}}(X_z) + e_{2,\mathrm{jpg}}(X_z)], \tag{30}$$

then

$$\boxed{\gamma_{s,g,\mathrm{jpg}} \geq g_{\mathrm{jpg},L}\gamma_{\mathrm{clean},L} - 4\overline{E}_{\mathrm{jpg},U} - \kappa_{\mathrm{jpg},U}.} \tag{31}$$

A concrete quality is certified when all constants refer to that exact recipe and the displayed right-hand side is positive.

Proof. Repeat the proof of Corollary B.5 with the exact endpoint bound $\epsilon_{\mathrm{jpg},U}$. The two-level non-commutation term pays both reconstruction levels for the exact practical map. □

For blur followed by down-up resizing, fix the kernel, boundary rule, antialias setting, reducing factor, output size, return size, and precision. Write the exact implementation as $T_{\mathrm{br}}u = Bu + r_{\mathrm{rnd}}(u)$, where B is the floating-point linear map before clipping and integer rounding. Let P_B project onto an alias-free subspace E_B, assume $b_{\min,L}\|v\|_N \le \|Bv\|_N \le b_{\max,U}\|v\|_N$ on E_B, and define $g_B = (b_{\min,L}^2 + b_{\max,U}^2)/2$ and $\epsilon_B = (b_{\max,U}^2 - b_{\min,L}^2)/2$. For $v_j = P_B F_j(x)$, set $\eta_{B,j}(x) = \|B(I-P_B)F_j(x) + r_{\mathrm{rnd}}(A^j x) - r_{\mathrm{rnd}}(A^{j+1}x)\|_N$ and

$$H_{B,j}(x) = \epsilon_B\|v_j\|_N^2 + g_B\|(I-P_B)F_j(x)\|_N^2 + 2b_{\max,U}\|v_j\|_N\eta_{B,j}(x) + \eta_{B,j}(x)^2. \tag{32}$$

Define $\delta_B = \max_{z\in\{s,g\}} \mathbb{E}_z[H_{B,0}(X_z) + H_{B,1}(X_z)]$.

Corollary B.7 (Practical blur–resize separation)**.** *Assume the preceding blur–resize conditions, the bounded invariant-domain and simultaneous clean-separation premises of Theorem 3.1, and a simultaneous bound*

$$\kappa_{B,U} \ge L_\rho \sum_{z\in\{s,g\}} \mathbb{E}_z[2k_{\mathrm{br}}(X_z) + e_{2,\mathrm{br}}(X_z)]. \tag{33}$$

Then

$$\boxed{\gamma_{s,g,\mathrm{br}} \ge g_B\gamma_{\mathrm{clean},L} - 2\delta_B - \kappa_{B,U}.} \tag{34}$$

The constants define the declared subspace, leakage envelope, rounding residual, and exact map.

Proof. On E_B, the squared gain differs from g_B by at most ϵ_B. Orthogonality gives $\|F_j\|_N^2 = \|v_j\|_N^2 + \|(I-P_B)F_j\|_N^2$. Expanding $\|Bv_j + w_j\|_N^2$ and using $\|Bv_j\|_N \le b_{\max,U}\|v_j\|_N$ proves the bound $H_{B,j}$. The two orbit edges give the classwise cost δ_B. Apply Theorem 3.1 with the exact distortion, $b_B = 0$, $\xi_B = 0$, and the displayed non-commutation cost. □

The specialization keeps codec and resize effects at the endpoint level because rounding and clipping are piecewise operations. A derivative-only approximation would omit cell transitions. Exact endpoint envelopes and hybrid-map telescoping preserve the practical recipe while retaining a readable separation inequality.

B.5 Additional Proofs for Shared Thresholds

For completeness, let $c = (s, g, T) \in \mathcal{C}$, let $F_{g,c}$ and $F_{s,c}$ be the population distribution functions of $q_1(TX_g)$ and $q_1(TX_s)$, and assign equality at the threshold to the real class. For $a \in [0,1]$, define $I_c^+(a) = \{\tau : 1 - F_{g,c}(\tau^-) \le a,\ F_{s,c}(\tau^-) \le a\}$ and $I_c^-(a) = \{\tau : F_{g,c}(\tau) \le a,\ 1 - F_{s,c}(\tau) \le a\}$.

Restatement (Theorem 3.2: Composite detection by one threshold)**.** A shared threshold rule whose two class errors are at most a in every declared cell exists exactly when

$$\bigcap_c I_c^+(a) \ne \varnothing \qquad \text{or} \qquad \bigcap_c I_c^-(a) \ne \varnothing. \tag{7}$$

Proof of Theorem 3.2. For orientation $+1$, generated error is $1 - F_{g,c}(\tau^-)$ and real error is $F_{s,c}(\tau^-)$. For orientation -1, the corresponding errors are $F_{g,c}(\tau)$ and $1 - F_{s,c}(\tau)$. Simultaneous feasibility is equivalent to one of the two displayed intersections. □

The full breakpoint class is used because any threshold between adjacent response values has identical empirical decisions. A heuristic threshold grid can create a false exclusion. The atom-aware use of $F(\tau^-)$ and $F(\tau)$ also fixes the assignment of ties and makes the criterion exact for discrete response laws.

B.6 Additional Response-Restricted Results and Condition Covers

For completeness, let P_c and Q_c be the real and generated laws of the scalar response in cell c. A randomized response decision is a measurable $\psi : \mathbb{R} \to [0,1]$, with balanced risk $R_c(\psi) = \frac{1}{2}\int \psi \, dP_c + \frac{1}{2}\int (1-\psi)\, dQ_c$. For a declared set $\mathcal{W}$ of probability vectors on the cell index, define $\overline{P}_w = \sum_c w_c P_c$ and $\overline{Q}_w = \sum_c w_c Q_c$.

Restatement (Theorem 3.3: Bayes response risk and least-favorable mixtures)**.** For every response decision ψ,

$$R_c(\psi) \geq \frac{1 - \mathrm{TV}(P_c, Q_c)}{2}, \tag{11}$$

whereas the Bayes optimum satisfies

$$R_c^* := \inf_{\psi} R_c(\psi) = \frac{1 - \mathrm{TV}(P_c, Q_c)}{2}. \tag{12}$$

Moreover,

$$\mathcal{R}_{q_1} := \inf_{\psi} \sup_{c} R_c(\psi) \geq \sup_{w \in \mathcal{W}} \frac{1 - \mathrm{TV}(\overline{P}_w, \overline{Q}_w)}{2}. \tag{13}$$

If an admissible w^* satisfies $\overline{P}_{w^*} = \overline{Q}_{w^*}$, then $\mathcal{R}_{q_1} = 1/2$.

Proof of Theorem 3.3. The simple-hypothesis identity is the standard variational characterization of total variation; it also gives the individual-rule inequality. For every w, $\sup_c R_c(\psi) \geq \sum_c w_c R_c(\psi)$. Taking the infimum over ψ and applying the identity to the two mixtures proves Eq. (13). Mixture equality gives the lower bound $1/2$, while any constant decision has balanced risk $1/2$ in every cell. □

For a fixed-capacity response-decision class $\mathcal{H}$, define $d_{\mathcal{H}}(c, c') = \sup_{h \in \mathcal{H}} |R_c(h) - R_{c'}(h)|$ and $\rho(J) = \sup_{c \in \mathcal{C}} \min_{j \in J} d_{\mathcal{H}}(c, j)$. Let $\widehat{h}_J$ be selected from calibration data using conditions in J. The radius premise may come from simultaneous samples over $\mathcal{C}$, a deterministic comparison $d_{\mathcal{H}} \leq L d_0$ in a known condition metric, or direct declaration of the evaluated index set. If data select J or $\widehat{h}_J$, the confidence event covers every candidate output.

Theorem B.8 (Condition-cover guarantee)**.** *Suppose that, on an event $\mathcal{E}_J$ of probability at least $1-\alpha$, simultaneous bounds satisfy $\widehat{\rho}_U(J) \geq \rho(J)$ and $\sup_{j \in J, h \in \mathcal{H}} |\widehat{R}_j(h) - R_j(h)| \leq \epsilon_U(J)$. Then*

$$\sup_{c \in \mathcal{C}} R_c(\widehat{h}_J) \leq \max_{j \in J} \widehat{R}_j(\widehat{h}_J) + \widehat{\rho}_U(J) + \epsilon_U(J). \tag{35}$$

Consequently the conservative form with $2\epsilon_U(J)$ also holds.

Proof. For each c, choose $j(c) \in J$ attaining, or arbitrarily approaching, the minimum in the definition of $\rho(J)$. Then

$$R_c(\widehat{h}_J) \leq R_{j(c)}(\widehat{h}_J) + d_{\mathcal{H}}(c, j(c)) \tag{36}$$

$$\leq \widehat{R}_{j(c)}(\widehat{h}_J) + \epsilon_U(J) + \widehat{\rho}_U(J). \tag{37}$$

Taking the supremum over c proves Eq. (35); the conservative form follows from $\epsilon_U(J) \geq 0$. □

For nested represented sets $J_1 \subset J_2 \subset \cdots$, hold the decision class, total data budget, and selection rule fixed. Let $R_J^* = \inf_{h \in \mathcal{H}} \max_{j \in J} R_j(h)$, $U(J) = R_J^* + \rho(J) + 2\epsilon(J)$, $B_k = \rho(J_k) - \rho(J_{k+1})$, and $C_k = [R_{J_{k+1}}^* - R_{J_k}^*] + 2[\epsilon(J_{k+1}) - \epsilon(J_k)]$.

Corollary B.9 (Benefit–conflict comparison)**.** *For every nested addition,*

$$U(J_{k+1}) - U(J_k) = C_k - B_k. \tag{38}$$

Thus $B_k > C_k$ decreases this certificate, $B_k < C_k$ increases it, and equality leaves it unchanged. If simultaneous intervals place $B_k - C_k$ strictly above or below zero, the certificate change has the opposite sign with the stated confidence. A statement about realized worst-cell risk additionally requires a confidence-valid change in that realized risk.

Proof. For any population minimax rule h_J^*, uniform convergence and empirical minimaxity give

$$\max_{j\in J} R_j(\widehat{h}_J) \le \max_{j\in J} \widehat{R}_j(\widehat{h}_J) + \epsilon(J) \tag{39}$$

$$\le \max_{j\in J} \widehat{R}_j(h_J^*) + \epsilon(J) \tag{40}$$

$$\le R_J^* + 2\epsilon(J). \tag{41}$$

The definition of $\rho(J)$ then gives the risk bound $\sup_c R_c(\widehat{h}_J) \le U(J)$. Subtract $U(J_k)$ from $U(J_{k+1})$ and collect the radius decrease into $-B_k$. Simultaneous interval arithmetic gives the confidence statement. □

Theorem B.10 (Missing response cluster)**.** *Suppose two possible worlds have identical laws for every observed source and differ only in one missing generated response cluster. On that cluster let $P = \frac{1}{2}(\delta_{-1} + \delta_{+1})$, $Q_0 = \delta_{-1}$, and $Q_1 = \delta_{+1}$. For any possibly randomized decision ψ measurable from q_1 and selected using observed-source data,*

$$\max\{R_0(\psi), R_1(\psi)\} \ge \frac{1}{2}. \tag{42}$$

The bound is attained by any constant response decision. Equivalently, the equal mixture of Q_0 and Q_1 equals P, so the construction matches Theorem 3.3 with zero slack.

Proof. Let $\mathcal{D}_{\mathrm{obs}}$ denote observed-source data and U collect rule-selection randomness. Conditional on $(\mathcal{D}_{\mathrm{obs}}, U)$, put $a = \psi(-1)$ and $b = \psi(+1)$. Direct calculation gives $R_0(\psi \mid \mathcal{D}_{\mathrm{obs}}, U) = \frac{1}{2} + (b-a)/4$ and $R_1(\psi \mid \mathcal{D}_{\mathrm{obs}}, U) = \frac{1}{2} + (a-b)/4$. Their sum equals one conditionally and after expectation, which yields the displayed maximum. □

The cover theorem supplies an upper certificate when represented conditions control the risk pseudometric. The missing-cluster construction supplies the matching limit when an observed response representation permits two indistinguishable worlds. These results define the statistical reach of q_1 decisions directly through risk geometry.

The response Bayes identity and the cover guarantee occupy opposite sides of the decision problem. Total variation quantifies the Bayes response risk from the full scalar law within one condition, independent of threshold shape. The shared-threshold class is narrower, so its empirical minimax risk can be higher. Condition coverage then transfers the risk of a selected response decision from represented conditions through $d_{\mathcal{H}}$. The radius depends on the decision class: a small radius for thresholds controls threshold rules, while arbitrary image classifiers require their own radius. This class dependence states exactly which observable differences must be estimated for the intended deployment rule.

The benefit–conflict identity further separates two effects of adding represented conditions. Coverage improves as $\rho(J)$ decreases, while the selected-set minimax risk and estimation cost can increase. Their difference determines the upper certificate. This decomposition suggests a principled source-selection objective: add a condition when its reduction in uncovered risk geometry exceeds the induced compatibility and estimation cost. The missing-cluster theorem shows the limiting case. If two unobserved generated clusters yield opposite point masses whose mixture equals the real response law, every response-measurable rule incurs balanced risk at least one half in one possible world.

B.7 Ordered Processing Boundaries

For a linear reconstruction let $R = I - A$ and $M = R^\top R - A^\top R^\top RA$, giving $q_1(x) = N^{-1}x^\top Mx$. Let $D = \mathbb{E}_g[XX^\top] - \mathbb{E}_s[XX^\top]$. For linear processing B, $\Delta_B = N^{-1}\operatorname{tr}(B^\top MBD)$. Consider a path $\{B_t : t \ge 0\}$ and a subspace E with orthogonal projector P_E. Let $u_1, \dots, u_m$ be an orthonormal unit basis of E that simultaneously diagonalizes the self-adjoint compressions $P_EMP_E\,|_E$ and $P_EDP_E\,|_E$. Assume that E is invariant under B_t and that $B_tu_j = e^{-\lambda_j t}u_j$ with $0 \le \lambda_1 < \cdots < \lambda_m$. With $a_j = N^{-1}(u_j^\top Mu_j)(u_j^\top Du_j)$,

$$\Gamma_E(t) = \sum_{j=1}^{m} a_j e^{-2\lambda_j t}. \tag{43}$$

The complete linear gap is $\Delta_{B_t} = \Gamma_E(t) + r_\perp(t)$, where $r_\perp$ contains the complement and cross-space terms.

Theorem B.11 (Single spectral crossing). *Suppose there is $k \in \{1, \dots, m-1\}$ such that $a_j \geq 0$ for $j \leq k$ and $a_j \leq 0$ for $j > k$, with at least one strict coefficient in each block. Then Γ_E has at most one zero on $[0, \infty)$. If $\Gamma_E(0) < 0$ and the coefficient associated with the smallest nonzero attenuation rate is positive, a unique $t_\star > 0$ exists and $\Gamma_E(t) < 0$ for $t < t_\star$, while $\Gamma_E(t) > 0$ for $t > t_\star$.*

Proof. After removing zero coefficients, write $P(t) = \sum_{j \leq k} a_j e^{-2\lambda_j t}$ and $Q(t) = \sum_{j>k} (-a_j) e^{-2\lambda_j t}$. Both functions are positive and $\Gamma_E = P - Q$. Their ratio satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t} \log \frac{Q(t)}{P(t)} = -2 \left(\mathbb{E}_{Q,t}[\lambda] - \mathbb{E}_{P,t}[\lambda] \right) < 0. \tag{44}$$

Every rate contributing to Q exceeds every rate contributing to P, so Q/P is strictly decreasing and $P = Q$ has at most one solution. The condition $\Gamma_E(0) < 0$ gives $Q(0)/P(0) > 1$. The smallest-rate nonzero positive term dominates as t grows, yielding $Q(t)/P(t) < 1$ for sufficiently large t. Continuity gives existence, and strict monotonicity gives uniqueness. □

The zero-count conclusion follows the generalized Descartes rule for ordered exponential sums (Voorhoeve, 1979). Total positivity and Pólya-frequency functions provide broader structural context for ordered kernels (Belton et al., 2023). For nonlinear reconstruction, write $\Delta(t) = \Gamma_E(t) + r(t)$, where r contains $r_\perp$, linearization error, remaining off-diagonal spectral terms, local curvature, and non-commutation.

Theorem B.12 (Stable processing boundary). *Let $I = [a, b]$ contain the unique root $t_\star$ of Γ_E. Suppose*

$$\inf_{t \in I} \Gamma_E'(t) \geq m > 0, \qquad \sup_{t \in I} |r'(t)| \leq \eta < m, \tag{45}$$

and

$$\Gamma_E(a) + \|r\|_{\infty, I} < 0, \qquad \Gamma_E(b) - \|r\|_{\infty, I} > 0. \tag{46}$$

Then Δ has exactly one root $\widehat{t}_\star \in (a, b)$, and

$$|\widehat{t}_\star - t_\star| \leq \frac{\|r\|_{\infty, I}}{m}. \tag{47}$$

Proof. Equation (45) gives $\Delta'(t) \geq m - \eta > 0$ on I, so Δ is strictly increasing. The endpoint margins give existence and uniqueness. Since $\Gamma_E(t_\star) = 0$ and $\Gamma_E(\widehat{t}_\star) = -r(\widehat{t}_\star)$, the mean-value theorem yields $m|\widehat{t}_\star - t_\star| \leq |r(\widehat{t}_\star)|$. □

For a discrete set of processing maps, an approximation $\widetilde{\Gamma}_T$ with $|\Delta_T - \widetilde{\Gamma}_T| \leq \varepsilon_T$ preserves the negative sign whenever $\widetilde{\Gamma}_T + \varepsilon_T < 0$. If centered class responses are sub-Gaussian with parameters $\sigma_s(t)$ and $\sigma_g(t)$ and $\delta(t) = -\Delta(t) > 0$, the midpoint threshold obeys

$$R_t \leq \frac{1}{2} \exp\left[-\frac{\delta(t)^2}{8\sigma_s(t)^2} \right] + \frac{1}{2} \exp\left[-\frac{\delta(t)^2}{8\sigma_g(t)^2} \right]. \tag{48}$$

Bounded-variable concentration can establish constituent mean deviations (Hoeffding, 1963). The response-restricted lower bound remains $(1 - \mathrm{TV}(P_{s,t}, P_{g,t}))/2$. The ordered construction is chosen because a monotone scalar processing strength can yield a unique interpretable boundary; the perturbation theorem makes every deviation from the shared-basis model an explicit remainder.

The simultaneous-diagonalization premise has a precise role. It removes cross terms with attenuation rates $\lambda_i + \lambda_j$ and leaves coefficients ordered by the single rates λ_j. A weaker orthogonality condition suffices when each cross contribution $(u_i^\top M u_j)(u_j^\top D u_i)$ vanishes. When cross terms remain, they are regrouped by their combined rates and their coefficient signs are checked again. The stable-boundary theorem then transfers the unique root from the ordered exponential sum to the nonlinear response gap through derivative dominance and endpoint margins. These conditions convert a qualitative observation of sign change into a located processing boundary with an explicit perturbation radius.

B.8 Assumption-to-Estimate Map

The clean-separation theorem combines paired geometric and statistical bounds. The operator-side quantities g_j, B_j, Q, and the derivative envelopes describe the fixed reconstruction map near the reference, while μ_z and M_z describe the local class distributions. A simultaneous population margin follows by applying favorable one-sided bounds to signed block contributions and unfavorable bounds to complement and remainder terms. The Taylor enclosure and chart-deviation cost remain explicit, so the fitted quadratic approximation is never treated as exact.

For deterministic processing, the most direct empirical object is the paired response decomposition in Eq. (2). Each image supplies its clean response, its transported reconstruction-edge response, and both non-commutation distances. The exact distortion option sets ξ_T to zero and estimates the centered distortion term from paired samples. A surrogate distortion can reduce computation when it comes with a separately bounded ξ_T. Because the center b_T is common to real and generated classes, it absorbs processing shifts shared by both populations while preserving their contrast. The lower bound then combines a clean source estimate with map-specific costs through one-sided interval arithmetic. A statement over several conditions uses simultaneous bounds across every condition index before taking the minimum processed margin.

For stochastic processing, the representative map and coupling determine the additional transport payment. A natural construction couples each channel output with the deterministic representative applied to the same input image. The expected coupled distance is then estimated separately within each class and multiplied by the response Lipschitz constant. Alternative couplings can tighten the bound when they are specified independently of confirmatory outcomes. The kernel remains common to the classes, which makes the comparison a property of how one channel acts on two input laws. Reporting the coupling-indexed quantity $L_K(T_0, \pi)$ keeps this dependence explicit and separates channel randomness from reconstruction–processing non-commutation.

The threshold theorems map directly to empirical distribution functions. Within every condition and class, all distinct response values define breakpoints, and left limits retain the correct atom assignment. Uniform CDF bands are evaluated at the full breakpoint union. Inner sets use the unfavorable band sides required to certify feasibility; outer sets use the favorable sides required to exclude it. Varying a traces the complete feasibility frontier, whose first nonempty level estimates E^*_{wc}. The same CDF event supplies lower and upper error bounds for every threshold, so minimizing the lower worst condition class error and balanced risk across the full class yields $E_{\mathrm{wc},L}$ and $R_{\mathrm{wc},L}$ independently of confirmatory data selection.

Condition coverage and spectral boundaries require different measurements. A cover analysis estimates pairwise risk changes under the fixed decision class or bounds them through a known condition metric; the radius is then the worst distance from an intended condition to its represented neighbor. The ordered spectral analysis estimates the coefficients, verifies their sign order, and bounds the complement, derivative, and endpoint remainders over an interval of processing strength. These two components answer distinct extrapolation questions. Coverage transfers decision risk across nearby conditions, while the spectral result locates a sign transition along an ordered path. Their assumptions remain separately testable and feed separate conclusions.

Together, these mappings define a reproducible sequence from response construction to a shared-rule conclusion. Geometry establishes a clean margin, paired processing measurements pay the survival costs, empirical distribution functions determine threshold compatibility, and risk geometry characterizes condition coverage. Every stage outputs a quantity used by the next theorem. A negative survival margin identifies insufficient retained response separation; empty threshold intersections identify distributional incompatibility; a large cover radius identifies sparse representation of the intended conditions. This localization turns an aggregate result into a specific mathematical quantity that can guide response redesign.

B.9 Stochastic Processing Extension

For a class-independent Markov kernel $K(x, \cdot)$, fix couplings π_z of (X_z, Y_z) with $Y_z \mid X_z \sim K(X_z, \cdot)$. Choose a deterministic representative T_0 satisfying Theorem 3.1. If q_1 is L_q-Lipschitz on the coupled support, define $\omega_K = L_q \sum_z \mathbb{E}_{\pi_z} \|Y_z - T_0 X_z\|_N$. The coarse value $L_q = L_\rho(1 + L_A)^2$ is valid under the preceding Lipschitz conditions.

Corollary B.13 (Stochastic separation after processing)**.** *Under the preceding coupling, suppose the signed deterministic enclosures in Theorem 3.1 hold for T_0. If $\omega_K \le \omega_U$, then*

$$\gamma_{s,g,K} \ge g_L \gamma_{\mathrm{clean},L} + d_L + c_L - \omega_U. \tag{49}$$

The endpoint specialization gives

$$\boxed{\gamma_{s,g,K} \ge L_K(T_0, \pi) := g_L \gamma_{\mathrm{clean},L} - 2\delta_U - \xi_U - \kappa_U - \omega_U.} \tag{50}$$

Each bound is indexed by the representative map and coupling.

Proof. Lipschitz continuity and the coupling give $|\mathbb{E}_z q_1(Y_z) - \mathbb{E}_z q_1(T_0 X_z)| \le L_q \mathbb{E}\|Y_z - T_0 X_z\|_N$. Applying this inequality to both class means subtracts at most ω_U from the deterministic signed bound or its endpoint specialization. □

The theorem offers two empirical levels. The signed contrast bound uses directly estimable $\Delta_{D,T}$ and $\Delta_{C,T}$ and preserves their direction. The endpoint specialization supplies a distribution free sufficient envelope through δ_U, ξ_U, and κ_U. Separating retained gain, signed contrasts, endpoint envelopes, and stochastic transport makes the result compatible with distinct deterministic maps, stochastic channels, and source-bound estimators.

Each term answers a different empirical question. The gain g_L measures how strongly the processing map retains the two clean reconstruction edges. The common center b_T removes a class-independent displacement of both edge responses, while δ_U measures the remaining classwise fluctuation around that center. The approximation term ξ_U is zero when the exact distortion is evaluated and is positive only when a tractable surrogate replaces it. The non-commutation term pays for applying reconstruction after processing instead of processing the clean reconstruction orbit. For a stochastic channel, ω_U adds the transport from a deterministic representative to coupled channel outputs. A positive lower bound therefore attributes survival to retained clean geometry after accounting for every observed loss.

B.10 Confidence-Valid Shared-Rule Bounds

Independent calibration data select $(\widehat{\sigma}, \widehat{\tau}) \in \arg\min_{\sigma,\tau} \max_c \widehat{R}_c^{\mathrm{cal}}(\sigma, \tau)$ over both orientations, all empirical breakpoints, and the two exterior constant rules. Ties are resolved lexicographically by point worst-cell balanced risk, point maximum false-positive rate over real-source-by-map projections, orientation order $+$ then $-$, and ascending threshold index from negative to positive infinity. The balanced risk is $R_c(\widehat{\sigma}, \widehat{\tau}) = \frac{1}{2}\mathrm{FPR}_c + \frac{1}{2}\mathrm{FNR}_c$. On an event $\mathcal{E}$ with probability at least $1 - \alpha$, suppose independent confirmatory data supply simultaneous risk and false-positive bounds as well as CDF bands $F_{z,c,L}(u) \le F_{z,c}(u) \le F_{z,c,U}(u)$. The Dvoretzky–Kiefer–Wolfowitz inequality supplies distribution-free uniform CDF control (Dvoretzky et al., 1956), with its sharp constant given in (Massart, 1990). Define

$$I^+_{c,\mathrm{in}}(a) = \{\tau : F_{g,c,L}(\tau^-) \ge 1 - a,\ F_{s,c,U}(\tau^-) \le a\}, \tag{51}$$

$$I^-_{c,\mathrm{in}}(a) = \{\tau : F_{g,c,U}(\tau) \le a,\ F_{s,c,L}(\tau) \ge 1 - a\}, \tag{52}$$

$$I^+_{c,\mathrm{out}}(a) = \{\tau : F_{g,c,U}(\tau^-) \ge 1 - a,\ F_{s,c,L}(\tau^-) \le a\}, \tag{53}$$

$$I^-_{c,\mathrm{out}}(a) = \{\tau : F_{g,c,L}(\tau) \le a,\ F_{s,c,U}(\tau) \ge 1 - a\}. \tag{54}$$

Theorem B.14 (Confidence-valid fixed-rule bounds)**.** *On $\mathcal{E}$,*

$$\sup_c R_c(\widehat{\sigma}, \widehat{\tau}) \le \max_c R_{c,U}, \qquad \sup_{s,T} \mathrm{FPR}_{s,T} \le \max_{s,T} P_{s,T,U}. \tag{55}$$

Moreover, at every level a, a nonempty intersection of the inner sets for either orientation certifies a population threshold meeting both class error constraints in every cell, while two empty outer intersections exclude such a threshold. If the same event supplies lower bounds $E_{\mathrm{wc},L}$ and $R_{\mathrm{wc},L}$ on E^*_{wc} and R^*_{wc}, then $E^*_{\mathrm{wc}} \ge E_{\mathrm{wc},L}$ and $R^*_{\mathrm{wc}} \ge R_{\mathrm{wc},L}$ on $\mathcal{E}$.

Proof. The first two conclusions are simultaneous upper confidence statements for a pair fixed independently of the confirmatory sample. Under the CDF bands, each inner set is contained in its population feasible set, and each population feasible set is contained in its outer set. These inclusions prove the intersection conclusions. The final claims follow from simultaneous lower confidence bounds for E^*_{wc} and R^*_{wc}. □

Table 3: Scale of the matched publicly available detector evaluation.

Dimension	Count	Construction
Publicly available methods	58	One canonical method and checkpoint per paper
Parent images	8,192	1,024 per source: two real and six generated
Processing conditions	32	Clean, Gaussian, blur–resize, and JPEG
Evaluation cells	384	2 real sources, 6 generators, 32 conditions
Scores per method	262,144	Every parent image under every condition
Detector scores	15,204,352	Complete matched evaluation across all methods
Controlled distributions	44,544	768 class-conditional distributions per method

C Additional Experimental Details

C.1 Implementation Details

All encoders and reconstruction operators are pretrained and fixed, so the confirmatory computation consists of scalar-response evaluation and statistical inference. Images are represented as canonical RGB arrays and processed by the fixed reconstruction implementation before q_1 is computed. The experimental sequence comprises small-scale validation, an implementation correctness check, sample-size analysis, a calibration split, a confirmatory split, and a controlled comparison. Calibration contains 803,104 response rows across 2,952 tasks; confirmatory evaluation contains 2,413,056 rows across 4,712 tasks. The calibration stage selects the orientation and threshold with a deterministic tie rule. Confirmatory responses provide all reported simultaneous bounds, preserving independence between rule selection and evaluation. Each empirical breakpoint is induced by an observed response value, and the exterior thresholds represent the two constant decisions. Tied response values define a single breakpoint, so repeated scores do not artificially enlarge the evaluated threshold class. This construction evaluates every distinct decision on the confirmatory responses. The same canonical byte-to-response transformation applies to both classes, and processing maps are class independent. Condition identifiers preserve real source, generator, and processing membership through aggregation, so worst-condition reduction is computed after cellwise errors. The controlled comparison uses fixed response definitions and source-linked invalidity states, so unavailable inputs remain unavailable endpoints throughout the analysis. Endpoint types remain distinct from response computation through simultaneous inference, keeping exact degeneracy separate from an estimated zero contrast. Consequently, an unavailable comparison is not interpreted as a measured zero effect or evidence supporting the corresponding scientific claim.

C.2 Additional Experimental Details

We evaluate 58 publicly available AI-generated image detection methods under one common decision protocol. The scope includes detectors whose original task classifies complete images produced by generative adversarial, diffusion, autoregressive, or commercial image generators. Face replacement, reenactment, identity editing, and localized manipulation methods fall outside this comparison. Each paper contributes one method and one canonical checkpoint. Checkpoint choice follows the authors' default inference release or primary cross-generator model and is fixed before response analysis. The resulting collection spans convolutional artifact detectors, local and frequency responses, vision foundation model adaptations, reconstruction methods, perturbation responses, and real-image reference models.

Every method produces one scalar score for the same 8,192 parents and 32 processing conditions, giving 262,144 scores per method and 15,204,352 scores in total. Table 3 summarizes the complete matched design. Author-released resizing, cropping, normalization, reconstruction, and score aggregation remain part of each method's response definition. Randomized image transformations receive a fixed seed determined by parent and condition identity. Per-image adaptation methods reset their state for every input. RealStats (Zisman & Shaham, 2026) and MDMF (Zhang et al., 2026) require a real-image reference distribution; both use the 512 clean real parents in calibration, with 256 examples from each real source, and freeze this state before confirmation. Their confirmatory responses access neither class labels nor other images in the evaluation batch.

Table 4: Source-specific response-atlas interval ranges.

Condition group	Rows	Min. lower	Max. upper	Invalid
Clean and Gaussian	156	-0.351837	0.352279	0
Blur and resize	156	-0.350522	0.350394	0
JPEG	72	-0.351650	0.352156	0

C.3 Additional Experimental Results

The original confirmatory responses give point critical shared class error $E^*_{\mathrm{wc}} = 0.993490$ with simultaneous lower confidence bound 0.906142. The point value means that every shared orientation and threshold misclassifies at least 99.35% of one class in some condition. The confidence bound preserves an error floor of 90.61% under simultaneous uncertainty.

Table 4 summarizes the source-bound response atlases. All 156 clean-and-Gaussian rows, all 156 deployment blur–resize rows, and all 72 JPEG rows are valid. Every atlas crosses zero. These intervals display substantial sign heterogeneity across conditions. Locally separated responses can point in different directions or occupy incompatible locations, raising the best attainable worst condition error of a fixed rule. The atlas locates which source and generator pairs change orientation or lose margin under a processing group, while the CDF analysis determines shared-rule feasibility. That structure guides the interpretation of transport and non-commutation terms in the theory.

The controlled comparison comprises 7,767 endpoints: 1,381 regular, 6,228 degenerate-exact, 98 unavailable-invalid-input, and 60 unavailable-full-range endpoints. The 7,609 eligible endpoints use 20,000 max-t replicates, order index 19,501, and critical value 3.957018 at probability 0.975. Table 5 keeps the endpoint types separate. Absolute q_1, e_0, encoder-based natural statistics, and the phase response all have their prespecified all-condition direction criteria refuted. The endpoint-state inventory records why many intervals are exact at zero and why unavailable rows remain outside numerical comparisons. Each response retains its original endpoint definition, and the simultaneous interval family tests the direction criterion shared by the comparisons. This organization distinguishes scientific refutation from numerical unavailability. A complete interval satisfies its family's registered support or refutation predicate; a target-crossing interval remains inconclusive unless a separate refutation condition holds, while an unavailable endpoint leaves that criterion unresolved.

D Additional Results and Analysis

D.1 Shared Rule Risk

Exhaustive evaluation of the q_1 decision class covers both orientations, every empirical breakpoint, and both exterior constant decisions. Table 1 gives a simultaneous lower bound of 0.456326 on minimax worst-condition balanced risk, so every shared threshold has an advantage over chance of at most 4.37 percentage points. The critical shared class error is 0.993490 with simultaneous lower bound 0.906142. Thus each shared rule has some condition in which one class incurs near-complete empirical error. The all-rule lower bounds and the selected-rule interval describe complementary parts of the same uncertainty geometry. Together they locate the attainable worst-condition performance without a user-selected success cutoff.

The minimizing high-fake orientation uses threshold 2.24737×10^{-4} and yields false-positive rate 0.993490 for the AIGI StyleGAN3 real-image source under JPEG quality 35. The low-fake orientation reaches a larger critical error of 0.996094, witnessed by Wukong generated images under JPEG qualities 35 and 50. Changing the common orientation therefore exchanges a severe real-side error for a severe generated-side error. These witnesses locate the threshold incompatibility in processed score distributions, while the complete threshold enumeration rules out numerical search as its explanation.

The selected operating point clarifies the mechanism. Calibration chooses the negative-infinity threshold in the positive orientation, producing a constant all-real decision. Its real-side false-positive interval is $[0, 0.087347]$, while its generated-side miss interval is $[0.912653, 1.0]$. A minimax calibration objective can favor such an endpoint when response locations shift across conditions: reducing an extreme false-positive cost in one part of the condition set can force generated errors

Table 5: Controlled comparison families and scientific outcomes.

Result family	Endpoints	Outcome
Repeated transmission and re-digitization	5	Supported
e_0 and absolute q_1	768	Refuted
Identity and bicubic responses	768	Refuted
ConvNeXt, DINOv2, DINOv3, and CLIP	3,072	Refuted; exact zero
Phase response	384	Refuted; exact zero
Real-source swap	32	Refuted
Canonical format and content parent	1,152	Inconclusive
Pooling-mask contrast	768	Refuted
Identity-conditioned oracle	384	Inconclusive
Condition-specific AUROC	192	Refuted
Compatibility, coverage, and overlap	124	Unavailable
Input validity and prospective adequacy	0	Refuted / inconclusive
Cover, addition, and spectral diagnostics	118	Inconclusive / refuted / unavailable

elsewhere. This reflects incompatibility across the evaluated conditionwise decision regions. It differs from a global disappearance of response information because several individual conditions still support useful ranking.

D.2 Additional Publicly Available Detector Results

MIRROR is strongest by all three shared-rule summaries: its calibration-selected worst-condition risk is 0.256, the complete threshold class reaches point risk 0.245, and its minimum worst class error is 0.483 with simultaneous lower bound 0.382. Across the collection, 53 simultaneous lower bounds on the critical class error exceed 0.5. Fig. 5 shows the separation between conditionwise ordering and one globally fixed decision rule.

D.3 Complete Confirmatory Response Data

The aggregate summaries above compress a large factorial evaluation. Table 6 defines the processing conditions. Tables 7, 8, and 9 expose every one of the 384 confirmatory response cells. Each cell uses 768 real and 768 generated responses and reports the estimated real-minus-generated difference together with its simultaneous interval. The entries remain separate by real-image source, generator, and processing condition, so the sign changes behind the shared-threshold result are directly visible. Displayed values are rounded to five decimals; the accompanying machine-readable table preserves full precision and source hashes.

Across the 384 point estimates, 256 are positive and 128 are negative. The split follows generator identity across both real-image sources and every processing condition: ADM, GLIDE, Midjourney, and StyleGAN3 contribute 64 positive cells each, while Stable Diffusion and Wukong contribute 64 negative cells each. This persistent orientation split supplies a direct empirical view of the response incompatibility that the atlas summarizes.

The largest point estimate occurs for the AIGC natural-image source, GLIDE, and the clean condition. The smallest occurs for the AIGI StyleGAN3 real-image source, Wukong, and the clean condition. All 384 simultaneous response intervals contain zero, so the tabled point-estimate pattern and the confidence-valid shared rule risk bound describe distinct layers of the result. The latter follows from the complete empirical CDF family and its exhaustive threshold reduction.

D.4 Controlled Comparison Families

Each comparison family has a prespecified two-sided decision rule. For all-cell direction claims, support requires every lower endpoint to exceed zero and refutation requires every upper endpoint to be nonpositive. The encoder-statistics family is supported when at least one statistic satisfies the all-cell lower-endpoint condition and refuted when every statistic satisfies the all-cell upper-endpoint condition. Equivalence controls use the band $(-0.00015, 0.00015)$; practical-band contrasts

Table 6: The complete processing-condition registry for the 384 confirmatory response cells.

Condition	Exact map ID	Setting
Clean	`clean`	Identity
Gaussian 1	`gaussian_t01`	$\sigma = 0.5774$ px
Gaussian 2	`gaussian_t02`	$\sigma = 0.8165$ px
Gaussian 3	`gaussian_t03`	$\sigma = 1.0000$ px
Gaussian 4	`gaussian_t04`	$\sigma = 1.1547$ px
Gaussian 5	`gaussian_t05`	$\sigma = 1.2910$ px
Gaussian 6	`gaussian_t06`	$\sigma = 1.4142$ px
Gaussian 7	`gaussian_t07`	$\sigma = 1.5275$ px
Gaussian 8	`gaussian_t08`	$\sigma = 1.6330$ px
Gaussian 9	`gaussian_t09`	$\sigma = 1.7321$ px
Gaussian 10	`gaussian_t10`	$\sigma = 1.8257$ px
Gaussian 11	`gaussian_t11`	$\sigma = 1.9149$ px
Gaussian 12	`gaussian_t12`	$\sigma = 2.0000$ px
Blur and resize 0	`deployment_t00`	$\sigma = 0.5000$ px, scale $= 0.750$
Blur and resize 1	`deployment_t01`	$\sigma = 0.6250$ px, scale $= 0.715$
Blur and resize 2	`deployment_t02`	$\sigma = 0.7500$ px, scale $= 0.680$
Blur and resize 3	`deployment_t03`	$\sigma = 0.8750$ px, scale $= 0.645$
Blur and resize 4	`deployment_t04`	$\sigma = 1.0000$ px, scale $= 0.610$
Blur and resize 5	`deployment_t05`	$\sigma = 1.1250$ px, scale $= 0.575$
Blur and resize 6	`deployment_t06`	$\sigma = 1.2500$ px, scale $= 0.540$
Blur and resize 7	`deployment_t07`	$\sigma = 1.3750$ px, scale $= 0.505$
Blur and resize 8	`deployment_t08`	$\sigma = 1.5000$ px, scale $= 0.470$
Blur and resize 9	`deployment_t09`	$\sigma = 1.6250$ px, scale $= 0.435$
Blur and resize 10	`deployment_t10`	$\sigma = 1.7500$ px, scale $= 0.400$
Blur and resize 11	`deployment_t11`	$\sigma = 1.8750$ px, scale $= 0.365$
Blur and resize 12	`deployment_t12`	$\sigma = 2.0000$ px, scale $= 0.330$
JPEG 95	`jpeg_q95`	quality $= 95$, 4:2:0
JPEG 85	`jpeg_q85`	quality $= 85$, 4:2:0
JPEG 75	`jpeg_q75`	quality $= 75$, 4:2:0
JPEG 65	`jpeg_q65`	quality $= 65$, 4:2:0
JPEG 50	`jpeg_q50`	quality $= 50$, 4:2:0
JPEG 35	`jpeg_q35`	quality $= 35$, 4:2:0

use $(-0.01, 0.01)$; and the oracle-gap support boundary is 0.02. An interval that crosses its target boundary is inconclusive unless the family-specific refutation condition also holds. Table 5 reports the resulting states without pooling their distinct scales.

The table partitions the 7,767 endpoints into named scientific families. Of these, 1,381 are regular, 6,228 are degenerate-exact, 98 are unavailable because their inputs are invalid, and 60 retain their full range. The unavailable compatibility comparison uses mismatched source-parent sets; the condition-coverage group contains the 60 full-range endpoints; and the zero-endpoint overlap comparison lacks a compatible empirical comparison. These states preserve missing information explicitly.

The prespecified direction criteria are refuted for absolute q_1, e_0, identity and bicubic responses, encoder-based natural statistics, and phase responses. Training-free reconstruction error remains a strong local detection mechanism (Ricker et al., 2024), which makes its all-condition outcome scientifically informative. ConvNeXt, DINOv2, DINOv3, and CLIP represent substantially different pretrained feature geometries. Each encoder family contributes 768 available endpoints whose lower and upper extrema are exactly zero. This exact-degenerate state refutes the positive directional criterion and leaves sign compatibility as a separate, unresolved mechanism.

The supported repeated-transmission and re-digitization result is a paired degradation statement. It resolves an auxiliary processing comparison under matched parents, while the main shared-threshold exclusion remains a distribution-level statement over all conditions. Keeping these estimands separate preserves the distinction between a paired mean contrast and threshold compatibility.

Table 7: Confirmatory q_1 cells for ADM and GLIDE. Each entry gives $\widehat{\gamma}\,[L, U]$ for the indicated real-image source.

Condition	ADM: AIGC natural	ADM: AIGI StyleGAN3
Clean	$0.00149\ [-0.34790, 0.35088]$	$0.00014\ [-0.34925, 0.34953]$
Gaussian 1	$0.00075\ [-0.34864, 0.35013]$	$0.00010\ [-0.34928, 0.34949]$
Gaussian 2	$0.00041\ [-0.34898, 0.34980]$	$0.00006\ [-0.34933, 0.34944]$
Gaussian 3	$0.00027\ [-0.34912, 0.34966]$	$0.00003\ [-0.34936, 0.34942]$
Gaussian 4	$0.00019\ [-0.34920, 0.34958]$	$0.00002\ [-0.34937, 0.34941]$
Gaussian 5	$0.00014\ [-0.34924, 0.34953]$	$0.00002\ [-0.34937, 0.34940]$
Gaussian 6	$0.00011\ [-0.34928, 0.34950]$	$0.00001\ [-0.34938, 0.34940]$
Gaussian 7	$0.00009\ [-0.34930, 0.34948]$	$0.00001\ [-0.34938, 0.34940]$
Gaussian 8	$0.00008\ [-0.34931, 0.34946]$	$0.00001\ [-0.34938, 0.34940]$
Gaussian 9	$0.00006\ [-0.34933, 0.34945]$	$0.00001\ [-0.34938, 0.34940]$
Gaussian 10	$0.00005\ [-0.34934, 0.34944]$	$0.00001\ [-0.34938, 0.34939]$
Gaussian 11	$0.00005\ [-0.34934, 0.34943]$	$0.00000\ [-0.34938, 0.34939]$
Gaussian 12	$0.00004\ [-0.34935, 0.34943]$	$0.00000\ [-0.34938, 0.34939]$
Blur and resize 0	$0.00061\ [-0.34878, 0.35000]$	$0.00008\ [-0.34930, 0.34947]$
Blur and resize 1	$0.00045\ [-0.34894, 0.34984]$	$0.00006\ [-0.34933, 0.34945]$
Blur and resize 2	$0.00034\ [-0.34905, 0.34973]$	$0.00004\ [-0.34935, 0.34943]$
Blur and resize 3	$0.00026\ [-0.34913, 0.34965]$	$0.00003\ [-0.34936, 0.34942]$
Blur and resize 4	$0.00020\ [-0.34919, 0.34959]$	$0.00002\ [-0.34937, 0.34941]$
Blur and resize 5	$0.00015\ [-0.34923, 0.34954]$	$0.00002\ [-0.34937, 0.34940]$
Blur and resize 6	$0.00012\ [-0.34927, 0.34951]$	$0.00001\ [-0.34938, 0.34940]$
Blur and resize 7	$0.00009\ [-0.34930, 0.34948]$	$0.00001\ [-0.34938, 0.34940]$
Blur and resize 8	$0.00007\ [-0.34932, 0.34946]$	$0.00001\ [-0.34938, 0.34940]$
Blur and resize 9	$0.00006\ [-0.34933, 0.34944]$	$0.00001\ [-0.34938, 0.34939]$
Blur and resize 10	$0.00004\ [-0.34935, 0.34943]$	$0.00000\ [-0.34938, 0.34939]$
Blur and resize 11	$0.00003\ [-0.34936, 0.34942]$	$0.00000\ [-0.34938, 0.34939]$
Blur and resize 12	$0.00003\ [-0.34936, 0.34942]$	$0.00000\ [-0.34938, 0.34939]$
JPEG 95	$0.00141\ [-0.34798, 0.35080]$	$0.00014\ [-0.34925, 0.34953]$
JPEG 85	$0.00142\ [-0.34797, 0.35081]$	$0.00012\ [-0.34927, 0.34951]$
JPEG 75	$0.00141\ [-0.34798, 0.35080]$	$0.00012\ [-0.34927, 0.34951]$
JPEG 65	$0.00140\ [-0.34799, 0.35078]$	$0.00012\ [-0.34927, 0.34951]$
JPEG 50	$0.00135\ [-0.34804, 0.35074]$	$0.00012\ [-0.34927, 0.34951]$
JPEG 35	$0.00129\ [-0.34810, 0.35068]$	$0.00011\ [-0.34927, 0.34950]$
Condition	**GLIDE: AIGC natural**	**GLIDE: AIGI StyleGAN3**
Clean	$0.00289\ [-0.34650, 0.35228]$	$0.00154\ [-0.34785, 0.35093]$
Gaussian 1	$0.00129\ [-0.34810, 0.35068]$	$0.00065\ [-0.34874, 0.35004]$
Gaussian 2	$0.00066\ [-0.34873, 0.35005]$	$0.00030\ [-0.34909, 0.34969]$
Gaussian 3	$0.00041\ [-0.34897, 0.34980]$	$0.00018\ [-0.34921, 0.34957]$
Gaussian 4	$0.00028\ [-0.34910, 0.34967]$	$0.00011\ [-0.34927, 0.34950]$
Gaussian 5	$0.00021\ [-0.34918, 0.34960]$	$0.00008\ [-0.34931, 0.34947]$
Gaussian 6	$0.00016\ [-0.34923, 0.34955]$	$0.00006\ [-0.34933, 0.34945]$
Gaussian 7	$0.00013\ [-0.34926, 0.34952]$	$0.00005\ [-0.34934, 0.34944]$
Gaussian 8	$0.00010\ [-0.34928, 0.34949]$	$0.00004\ [-0.34935, 0.34943]$
Gaussian 9	$0.00009\ [-0.34930, 0.34947]$	$0.00003\ [-0.34936, 0.34942]$
Gaussian 10	$0.00007\ [-0.34932, 0.34946]$	$0.00002\ [-0.34937, 0.34941]$
Gaussian 11	$0.00006\ [-0.34933, 0.34945]$	$0.00002\ [-0.34937, 0.34941]$
Gaussian 12	$0.00005\ [-0.34934, 0.34944]$	$0.00001\ [-0.34937, 0.34940]$
Blur and resize 0	$0.00101\ [-0.34838, 0.35039]$	$0.00048\ [-0.34891, 0.34987]$
Blur and resize 1	$0.00072\ [-0.34867, 0.35011]$	$0.00033\ [-0.34906, 0.34972]$
Blur and resize 2	$0.00053\ [-0.34885, 0.34992]$	$0.00023\ [-0.34915, 0.34962]$
Blur and resize 3	$0.00040\ [-0.34899, 0.34979]$	$0.00017\ [-0.34922, 0.34955]$
Blur and resize 4	$0.00030\ [-0.34909, 0.34969]$	$0.00012\ [-0.34927, 0.34951]$
Blur and resize 5	$0.00022\ [-0.34917, 0.34961]$	$0.00008\ [-0.34930, 0.34947]$
Blur and resize 6	$0.00017\ [-0.34922, 0.34956]$	$0.00006\ [-0.34933, 0.34945]$
Blur and resize 7	$0.00013\ [-0.34926, 0.34952]$	$0.00004\ [-0.34934, 0.34943]$
Blur and resize 8	$0.00010\ [-0.34929, 0.34948]$	$0.00003\ [-0.34936, 0.34942]$
Blur and resize 9	$0.00007\ [-0.34932, 0.34946]$	$0.00002\ [-0.34937, 0.34941]$
Blur and resize 10	$0.00005\ [-0.34933, 0.34944]$	$0.00002\ [-0.34937, 0.34940]$
Blur and resize 11	$0.00004\ [-0.34935, 0.34943]$	$0.00001\ [-0.34938, 0.34940]$
Blur and resize 12	$0.00003\ [-0.34936, 0.34942]$	$0.00001\ [-0.34938, 0.34940]$
JPEG 95	$0.00274\ [-0.34665, 0.35213]$	$0.00147\ [-0.34792, 0.35086]$
JPEG 85	$0.00277\ [-0.34662, 0.35216]$	$0.00147\ [-0.34792, 0.35086]$
JPEG 75	$0.00273\ [-0.34666, 0.35212]$	$0.00144\ [-0.34795, 0.35083]$
JPEG 65	$0.00267\ [-0.34672, 0.35206]$	$0.00140\ [-0.34799, 0.35079]$
JPEG 50	$0.00256\ [-0.34682, 0.35195]$	$0.00134\ [-0.34805, 0.35073]$
JPEG 35	$0.00243\ [-0.34696, 0.35182]$	$0.00126\ [-0.34813, 0.35064]$

Table 8: Confirmatory q_1 cells for Midjourney and Stable Diffusion. Each entry gives $\widehat{\gamma}\,[L, U]$ for the indicated real-image source.

Condition	Midjourney: AIGC natural	Midjourney: AIGI StyleGAN3
Clean	$0.00181\ [-0.34758, 0.35120]$	$0.00046\ [-0.34893, 0.34985]$
Gaussian 1	$0.00084\ [-0.34855, 0.35023]$	$0.00020\ [-0.34919, 0.34959]$
Gaussian 2	$0.00046\ [-0.34893, 0.34985]$	$0.00010\ [-0.34929, 0.34949]$
Gaussian 3	$0.00030\ [-0.34909, 0.34969]$	$0.00006\ [-0.34933, 0.34945]$
Gaussian 4	$0.00021\ [-0.34918, 0.34960]$	$0.00004\ [-0.34935, 0.34943]$
Gaussian 5	$0.00016\ [-0.34923, 0.34955]$	$0.00003\ [-0.34936, 0.34942]$
Gaussian 6	$0.00012\ [-0.34926, 0.34951]$	$0.00002\ [-0.34937, 0.34941]$
Gaussian 7	$0.00010\ [-0.34929, 0.34949]$	$0.00002\ [-0.34937, 0.34941]$
Gaussian 8	$0.00008\ [-0.34931, 0.34947]$	$0.00001\ [-0.34937, 0.34940]$
Gaussian 9	$0.00007\ [-0.34932, 0.34946]$	$0.00001\ [-0.34938, 0.34940]$
Gaussian 10	$0.00006\ [-0.34933, 0.34945]$	$0.00001\ [-0.34938, 0.34940]$
Gaussian 11	$0.00005\ [-0.34934, 0.34944]$	$0.00001\ [-0.34938, 0.34940]$
Gaussian 12	$0.00004\ [-0.34935, 0.34943]$	$0.00001\ [-0.34938, 0.34940]$
Blur and resize 0	$0.00068\ [-0.34871, 0.35006]$	$0.00015\ [-0.34924, 0.34954]$
Blur and resize 1	$0.00050\ [-0.34889, 0.34989]$	$0.00011\ [-0.34928, 0.34950]$
Blur and resize 2	$0.00038\ [-0.34901, 0.34977]$	$0.00008\ [-0.34931, 0.34947]$
Blur and resize 3	$0.00029\ [-0.34909, 0.34968]$	$0.00006\ [-0.34933, 0.34945]$
Blur and resize 4	$0.00022\ [-0.34916, 0.34961]$	$0.00004\ [-0.34934, 0.34943]$
Blur and resize 5	$0.00017\ [-0.34922, 0.34956]$	$0.00003\ [-0.34936, 0.34942]$
Blur and resize 6	$0.00013\ [-0.34926, 0.34952]$	$0.00002\ [-0.34936, 0.34941]$
Blur and resize 7	$0.00010\ [-0.34929, 0.34949]$	$0.00002\ [-0.34937, 0.34941]$
Blur and resize 8	$0.00008\ [-0.34931, 0.34947]$	$0.00001\ [-0.34938, 0.34940]$
Blur and resize 9	$0.00006\ [-0.34933, 0.34945]$	$0.00001\ [-0.34938, 0.34940]$
Blur and resize 10	$0.00005\ [-0.34934, 0.34943]$	$0.00001\ [-0.34938, 0.34940]$
Blur and resize 11	$0.00003\ [-0.34935, 0.34942]$	$0.00000\ [-0.34938, 0.34939]$
Blur and resize 12	$0.00003\ [-0.34936, 0.34941]$	$0.00000\ [-0.34939, 0.34939]$
JPEG 95	$0.00168\ [-0.34771, 0.35106]$	$0.00040\ [-0.34899, 0.34979]$
JPEG 85	$0.00169\ [-0.34770, 0.35108]$	$0.00039\ [-0.34900, 0.34978]$
JPEG 75	$0.00167\ [-0.34772, 0.35106]$	$0.00038\ [-0.34901, 0.34977]$
JPEG 65	$0.00163\ [-0.34775, 0.35102]$	$0.00036\ [-0.34903, 0.34975]$
JPEG 50	$0.00156\ [-0.34783, 0.35095]$	$0.00034\ [-0.34905, 0.34972]$
JPEG 35	$0.00147\ [-0.34792, 0.35086]$	$0.00030\ [-0.34909, 0.34969]$
Condition	**SD 1.4: AIGC natural**	**SD 1.4: AIGI StyleGAN3**
Clean	$-0.00051\ [-0.34990, 0.34888]$	$-0.00186\ [-0.35125, 0.34753]$
Gaussian 1	$-0.00035\ [-0.34974, 0.34904]$	$-0.00099\ [-0.35038, 0.34840]$
Gaussian 2	$-0.00023\ [-0.34962, 0.34916]$	$-0.00059\ [-0.34997, 0.34880]$
Gaussian 3	$-0.00017\ [-0.34956, 0.34922]$	$-0.00040\ [-0.34979, 0.34899]$
Gaussian 4	$-0.00013\ [-0.34952, 0.34926]$	$-0.00030\ [-0.34969, 0.34909]$
Gaussian 5	$-0.00010\ [-0.34949, 0.34929]$	$-0.00023\ [-0.34962, 0.34916]$
Gaussian 6	$-0.00008\ [-0.34947, 0.34931]$	$-0.00018\ [-0.34957, 0.34921]$
Gaussian 7	$-0.00007\ [-0.34946, 0.34932]$	$-0.00015\ [-0.34954, 0.34924]$
Gaussian 8	$-0.00006\ [-0.34945, 0.34933]$	$-0.00012\ [-0.34951, 0.34926]$
Gaussian 9	$-0.00005\ [-0.34944, 0.34934]$	$-0.00010\ [-0.34949, 0.34928]$
Gaussian 10	$-0.00004\ [-0.34943, 0.34935]$	$-0.00009\ [-0.34948, 0.34930]$
Gaussian 11	$-0.00004\ [-0.34943, 0.34935]$	$-0.00008\ [-0.34947, 0.34931]$
Gaussian 12	$-0.00003\ [-0.34942, 0.34936]$	$-0.00007\ [-0.34946, 0.34932]$
Blur and resize 0	$-0.00032\ [-0.34971, 0.34907]$	$-0.00085\ [-0.35023, 0.34854]$
Blur and resize 1	$-0.00026\ [-0.34964, 0.34913]$	$-0.00065\ [-0.35004, 0.34874]$
Blur and resize 2	$-0.00021\ [-0.34960, 0.34918]$	$-0.00051\ [-0.34990, 0.34888]$
Blur and resize 3	$-0.00017\ [-0.34955, 0.34922]$	$-0.00040\ [-0.34979, 0.34899]$
Blur and resize 4	$-0.00013\ [-0.34952, 0.34925]$	$-0.00031\ [-0.34970, 0.34908]$
Blur and resize 5	$-0.00011\ [-0.34950, 0.34928]$	$-0.00025\ [-0.34963, 0.34914]$
Blur and resize 6	$-0.00009\ [-0.34947, 0.34930]$	$-0.00019\ [-0.34958, 0.34920]$
Blur and resize 7	$-0.00007\ [-0.34946, 0.34932]$	$-0.00015\ [-0.34954, 0.34924]$
Blur and resize 8	$-0.00005\ [-0.34944, 0.34933]$	$-0.00012\ [-0.34951, 0.34927]$
Blur and resize 9	$-0.00004\ [-0.34943, 0.34934]$	$-0.00009\ [-0.34948, 0.34929]$
Blur and resize 10	$-0.00004\ [-0.34943, 0.34935]$	$-0.00008\ [-0.34946, 0.34931]$
Blur and resize 11	$-0.00003\ [-0.34942, 0.34936]$	$-0.00006\ [-0.34945, 0.34933]$
Blur and resize 12	$-0.00003\ [-0.34941, 0.34936]$	$-0.00005\ [-0.34944, 0.34934]$
JPEG 95	$-0.00054\ [-0.34993, 0.34884]$	$-0.00182\ [-0.35121, 0.34757]$
JPEG 85	$-0.00054\ [-0.34992, 0.34885]$	$-0.00184\ [-0.35122, 0.34755]$
JPEG 75	$-0.00053\ [-0.34992, 0.34886]$	$-0.00182\ [-0.35121, 0.34756]$
JPEG 65	$-0.00054\ [-0.34993, 0.34885]$	$-0.00181\ [-0.35120, 0.34758]$
JPEG 50	$-0.00055\ [-0.34994, 0.34884]$	$-0.00177\ [-0.35116, 0.34761]$
JPEG 35	$-0.00055\ [-0.34994, 0.34884]$	$-0.00173\ [-0.35111, 0.34766]$

Table 9: Confirmatory q_1 cells for StyleGAN3 and Wukong. Each entry gives $\widehat{\gamma}\,[L, U]$ for the indicated real-image source.

Condition	**StyleGAN3: AIGC natural**	**StyleGAN3: AIGI StyleGAN3**
Clean	$0.00212\ [-0.34727, 0.35151]$	$0.00077\ [-0.34862, 0.35015]$
Gaussian 1	$0.00101\ [-0.34838, 0.35040]$	$0.00037\ [-0.34902, 0.34976]$
Gaussian 2	$0.00056\ [-0.34883, 0.34995]$	$0.00020\ [-0.34918, 0.34959]$
Gaussian 3	$0.00038\ [-0.34901, 0.34977]$	$0.00014\ [-0.34925, 0.34953]$
Gaussian 4	$0.00027\ [-0.34912, 0.34966]$	$0.00010\ [-0.34929, 0.34949]$
Gaussian 5	$0.00021\ [-0.34918, 0.34960]$	$0.00008\ [-0.34931, 0.34947]$
Gaussian 6	$0.00017\ [-0.34922, 0.34955]$	$0.00006\ [-0.34932, 0.34945]$
Gaussian 7	$0.00014\ [-0.34925, 0.34953]$	$0.00005\ [-0.34933, 0.34944]$
Gaussian 8	$0.00012\ [-0.34927, 0.34950]$	$0.00005\ [-0.34934, 0.34944]$
Gaussian 9	$0.00010\ [-0.34929, 0.34949]$	$0.00004\ [-0.34935, 0.34943]$
Gaussian 10	$0.00008\ [-0.34930, 0.34947]$	$0.00004\ [-0.34935, 0.34942]$
Gaussian 11	$0.00007\ [-0.34932, 0.34946]$	$0.00003\ [-0.34936, 0.34942]$
Gaussian 12	$0.00006\ [-0.34933, 0.34945]$	$0.00003\ [-0.34936, 0.34942]$
Blur and resize 0	$0.00082\ [-0.34857, 0.35021]$	$0.00030\ [-0.34909, 0.34968]$
Blur and resize 1	$0.00061\ [-0.34878, 0.35000]$	$0.00022\ [-0.34917, 0.34961]$
Blur and resize 2	$0.00047\ [-0.34892, 0.34986]$	$0.00017\ [-0.34922, 0.34956]$
Blur and resize 3	$0.00037\ [-0.34902, 0.34976]$	$0.00013\ [-0.34925, 0.34952]$
Blur and resize 4	$0.00029\ [-0.34910, 0.34967]$	$0.00011\ [-0.34928, 0.34949]$
Blur and resize 5	$0.00022\ [-0.34917, 0.34961]$	$0.00008\ [-0.34930, 0.34947]$
Blur and resize 6	$0.00017\ [-0.34921, 0.34956]$	$0.00007\ [-0.34932, 0.34946]$
Blur and resize 7	$0.00014\ [-0.34925, 0.34953]$	$0.00006\ [-0.34933, 0.34944]$
Blur and resize 8	$0.00011\ [-0.34928, 0.34950]$	$0.00005\ [-0.34934, 0.34943]$
Blur and resize 9	$0.00009\ [-0.34930, 0.34948]$	$0.00004\ [-0.34935, 0.34943]$
Blur and resize 10	$0.00007\ [-0.34932, 0.34946]$	$0.00003\ [-0.34936, 0.34942]$
Blur and resize 11	$0.00006\ [-0.34933, 0.34944]$	$0.00003\ [-0.34936, 0.34941]$
Blur and resize 12	$0.00004\ [-0.34934, 0.34943]$	$0.00002\ [-0.34937, 0.34941]$
JPEG 95	$0.00194\ [-0.34745, 0.35133]$	$0.00067\ [-0.34872, 0.35006]$
JPEG 85	$0.00199\ [-0.34740, 0.35138]$	$0.00069\ [-0.34870, 0.35008]$
JPEG 75	$0.00200\ [-0.34738, 0.35139]$	$0.00071\ [-0.34867, 0.35010]$
JPEG 65	$0.00199\ [-0.34740, 0.35138]$	$0.00072\ [-0.34867, 0.35011]$
JPEG 50	$0.00194\ [-0.34745, 0.35133]$	$0.00072\ [-0.34867, 0.35010]$
JPEG 35	$0.00187\ [-0.34751, 0.35126]$	$0.00070\ [-0.34869, 0.35009]$
Condition	**Wukong: AIGC natural**	**Wukong: AIGI StyleGAN3**
Clean	$-0.00110\ [-0.35049, 0.34829]$	$-0.00245\ [-0.35184, 0.34694]$
Gaussian 1	$-0.00068\ [-0.35006, 0.34871]$	$-0.00132\ [-0.35071, 0.34807]$
Gaussian 2	$-0.00044\ [-0.34983, 0.34895]$	$-0.00080\ [-0.35019, 0.34859]$
Gaussian 3	$-0.00033\ [-0.34972, 0.34906]$	$-0.00056\ [-0.34995, 0.34882]$
Gaussian 4	$-0.00026\ [-0.34965, 0.34913]$	$-0.00043\ [-0.34982, 0.34896]$
Gaussian 5	$-0.00021\ [-0.34960, 0.34918]$	$-0.00034\ [-0.34973, 0.34905]$
Gaussian 6	$-0.00018\ [-0.34956, 0.34921]$	$-0.00028\ [-0.34967, 0.34911]$
Gaussian 7	$-0.00015\ [-0.34954, 0.34924]$	$-0.00023\ [-0.34962, 0.34916]$
Gaussian 8	$-0.00013\ [-0.34952, 0.34926]$	$-0.00020\ [-0.34959, 0.34919]$
Gaussian 9	$-0.00011\ [-0.34950, 0.34928]$	$-0.00017\ [-0.34956, 0.34922]$
Gaussian 10	$-0.00010\ [-0.34949, 0.34929]$	$-0.00015\ [-0.34954, 0.34924]$
Gaussian 11	$-0.00009\ [-0.34948, 0.34930]$	$-0.00013\ [-0.34952, 0.34926]$
Gaussian 12	$-0.00008\ [-0.34947, 0.34931]$	$-0.00011\ [-0.34950, 0.34928]$
Blur and resize 0	$-0.00061\ [-0.35000, 0.34878]$	$-0.00113\ [-0.35052, 0.34826]$
Blur and resize 1	$-0.00049\ [-0.34988, 0.34890]$	$-0.00088\ [-0.35027, 0.34851]$
Blur and resize 2	$-0.00040\ [-0.34979, 0.34899]$	$-0.00070\ [-0.35009, 0.34869]$
Blur and resize 3	$-0.00033\ [-0.34972, 0.34906]$	$-0.00056\ [-0.34995, 0.34883]$
Blur and resize 4	$-0.00027\ [-0.34966, 0.34912]$	$-0.00045\ [-0.34984, 0.34894]$
Blur and resize 5	$-0.00022\ [-0.34961, 0.34917]$	$-0.00036\ [-0.34975, 0.34903]$
Blur and resize 6	$-0.00018\ [-0.34957, 0.34920]$	$-0.00029\ [-0.34968, 0.34910]$
Blur and resize 7	$-0.00015\ [-0.34954, 0.34924]$	$-0.00023\ [-0.34962, 0.34915]$
Blur and resize 8	$-0.00013\ [-0.34951, 0.34926]$	$-0.00019\ [-0.34958, 0.34920]$
Blur and resize 9	$-0.00010\ [-0.34949, 0.34928]$	$-0.00015\ [-0.34954, 0.34923]$
Blur and resize 10	$-0.00009\ [-0.34948, 0.34930]$	$-0.00013\ [-0.34951, 0.34926]$
Blur and resize 11	$-0.00007\ [-0.34946, 0.34932]$	$-0.00010\ [-0.34949, 0.34929]$
Blur and resize 12	$-0.00007\ [-0.34945, 0.34932]$	$-0.00009\ [-0.34948, 0.34930]$
JPEG 95	$-0.00096\ [-0.35035, 0.34843]$	$-0.00223\ [-0.35162, 0.34716]$
JPEG 85	$-0.00096\ [-0.35035, 0.34843]$	$-0.00226\ [-0.35165, 0.34713]$
JPEG 75	$-0.00097\ [-0.35036, 0.34842]$	$-0.00226\ [-0.35165, 0.34713]$
JPEG 65	$-0.00098\ [-0.35037, 0.34840]$	$-0.00226\ [-0.35165, 0.34713]$
JPEG 50	$-0.00101\ [-0.35039, 0.34838]$	$-0.00223\ [-0.35162, 0.34716]$
JPEG 35	$-0.00102\ [-0.35041, 0.34837]$	$-0.00220\ [-0.35159, 0.34719]$

The max-t construction ties the 7,609 eligible endpoints to one multiplicity-controlled comparison family. Its 20,000 replicates use order index 19,501 and critical value 3.957018 at probability 0.975. A common resampling reference makes simultaneous interval interpretation possible while preserving endpoint-specific scales. The degenerate-exact endpoints enter with exact behavior, while unavailable endpoints contribute no numerical value.

D.5 Mechanism of Threshold Incompatibility

The principal empirical failure mode is source–condition incompatibility for a single scalar threshold. The selected constant all-real rule incurs a generated-side miss lower bound of 0.912653 even though its real-side false-positive upper bound is 0.087347. This asymmetric error profile shows how minimax selection responds to heterogeneous response locations. The risk is spread through a conservative endpoint decision whose errors are easy to characterize but operationally weak for generated-image discovery.

The response atlas explains why threshold adjustment leaves the conflict intact. Cellwise intervals occupy both signs across each major processing group, while condition-specific rankings vary from strongly reversed to moderately useful. Moving the threshold within one orientation trades error between incompatible cells; flipping the common orientation exchanges which subset is favored. The exhaustive threshold result measures the smallest remaining worst condition error after both adjustments. A response with stable sign and compatible class locations across sources and generators would address the mechanism directly. The theory suggests measurable design criteria for such a response: a positive clean separation lower bound, small centered transport and non-commutation costs, and a small critical shared class error. A candidate can be screened at each layer before full confirmatory evaluation. A weak mean-survival bound points to response construction or processing sensitivity, while a large critical error points to scale or tail alignment. This diagnosis is actionable because response magnitude and location can be redesigned separately: the operator can target a stable change statistic, and calibration can test the resulting critical error.

Secondary endpoint states distinguish resolved mechanisms from open ones. Full-range endpoints convey no directional resolution, and incompatible parent sets prevent a shared comparison. Matched source construction with a prespecified comparison estimand would resolve these entries through simultaneous intervals on compatible endpoints. The primary shared rule risk bound follows from the independent CDF family; the secondary inventory locates the remaining mechanistic uncertainty.

D.6 Additional Response-Atlas Interpretation

The response atlases in Table 4 are complete and valid for clean-and-Gaussian, deployment blur–resize, and JPEG groups. Their interval ranges cross zero in every group. This pattern means that a fixed orientation requires cellwise signed evidence; the average response contrast alone is insufficient. It also explains why the processing-survival theorem separates the clean lower bound from transport and non-commutation costs. A positive source-bound lower bound must hold for the relevant pair before a processing specialization can retain it; averaging positive and negative cells would erase the premise needed by the theorem.

The signed AUROC range also prevents a misleading interpretation of low values as complete absence of signal. Reversing an orientation can turn a low local ranking into a high one, but the orientation is shared across $\mathcal{C}$. Opposite local orientations therefore create conflict. The exact threshold-intersection theorem formalizes this point at the distribution level, while the critical shared class error measures the resulting conflict after simultaneous uncertainty is included.

D.7 Additional Public Detector Interpretation

The publicly available detector comparison evaluates whether the shared threshold phenomenon extends beyond the reconstruction response developed in this paper. The 58 methods span convolutional artifact detectors, local and frequency responses, vision foundation model adaptations, reconstruction methods, perturbation responses, and real image reference models. The exact intersection and minimax risk results apply to every scalar score, so this experiment instantiates the general decision theory with released detector responses while the processing decomposition remains an analysis of q_1.

Table 10: Complete 58-method shared-threshold comparison. The first four numeric columns report mean cellwise orientation-free AUROC (Gauss.: Gaussian; Blur: blur–resize). Fixed and Best report calibrated and minimum attainable balanced risk; $E^\star$ is minimum worst-class error, with its lower bound in brackets.

Method	**Clean**	**Gauss.**	**Blur**	**JPEG**	**Fixed**	**Best**	$E^\star$ **[lower]**
AEROBLADE (Ricker et al., 2024)	0.682	0.649	0.649	0.657	0.500	0.500	1.000 [0.899]
AIDE (Yan et al., 2025b)	0.771	0.684	0.687	0.683	0.500	0.500	0.932 [0.831]
AlignedForensics (Rajan et al., 2025)	0.826	0.795	0.773	0.800	0.500	0.500	0.891 [0.789]
B-Free (Guillaro et al., 2025)	0.905	0.872	0.854	0.849	0.387	0.382	0.484 [0.383]
C2P-CLIP (Tan et al., 2025)	0.887	0.766	0.770	0.636	0.500	0.500	0.953 [0.852]
CLIDE (Betser et al., 2026)	0.556	0.582	0.585	0.589	0.500	0.500	0.868 [0.767]
CLIPping (Khan & Dang-Nguyen, 2024)	0.701	0.629	0.629	0.629	0.500	0.500	0.951 [0.849]
CNNSpot (Wang et al., 2020)	0.672	0.639	0.636	0.640	0.500	0.500	0.697 [0.595]
CO-SPY (Cheng et al., 2025)	0.917	0.606	0.611	0.605	0.500	0.500	0.986 [0.884]
Community Forensics (Park & Owens, 2025)	0.991	0.938	0.933	0.934	0.492	0.492	0.969 [0.867]
ConV (Zhang et al., 2025c)	0.643	0.637	0.642	0.644	0.500	0.500	0.914 [0.813]
Corvi detector (Corvi et al., 2023)	0.833	0.766	0.728	0.732	0.500	0.500	0.884 [0.783]
D3 (Yang et al., 2025)	0.970	0.929	0.926	0.931	0.380	0.364	0.516 [0.414]
D3QE (Zhang et al., 2025b)	0.861	0.781	0.784	0.780	0.467	0.462	0.570 [0.469]
DE-FAKE (Sha et al., 2023)	0.819	0.743	0.744	0.754	0.500	0.500	0.702 [0.600]
DEAR (Kim et al., 2026)	0.842	0.779	0.736	0.746	0.500	0.500	0.707 [0.606]
DeeCLIP (Keita et al., 2025)	0.758	0.659	0.655	0.597	0.500	0.500	0.668 [0.567]
DGS-Net (Yan et al., 2025a)	0.932	0.717	0.715	0.610	0.500	0.500	0.999 [0.897]
DIRE (Wang et al., 2023)	0.530	0.524	0.528	0.560	0.500	0.500	0.947 [0.845]
DRCT (Chen et al., 2024a)	0.893	0.768	0.767	0.789	0.472	0.466	0.783 [0.681]
Dual Data Alignment (Chen et al., 2025a)	0.958	0.873	0.879	0.856	0.475	0.450	0.720 [0.619]
FatFormer (Liu et al., 2024)	0.820	0.680	0.686	0.590	0.500	0.500	0.953 [0.852]
FerretNet (Liang et al., 2025)	0.932	0.636	0.629	0.578	0.500	0.500	0.988 [0.887]
FIRE (Chu et al., 2025)	0.569	0.542	0.540	0.591	0.500	0.500	1.000 [0.899]
Forensic-MoE (Fang et al., 2025)	0.883	0.708	0.708	0.669	0.500	0.500	0.975 [0.874]
ForgeLens (Chen et al., 2025b)	0.926	0.732	0.741	0.728	0.500	0.500	0.997 [0.896]
FreqDetect (Frank et al., 2020)	0.645	0.657	0.628	0.587	0.500	0.500	0.975 [0.874]
FreqNet (Tan et al., 2024b)	0.795	0.560	0.568	0.589	0.500	0.500	0.984 [0.883]
FSD (Nguyen et al., 2025)	0.557	0.552	0.569	0.539	0.500	0.500	0.927 [0.826]
Fusing (Ju et al., 2022)	0.662	0.619	0.618	0.634	0.500	0.500	0.883 [0.781]
GAPL (Qin et al., 2026)	0.996	0.981	0.978	0.972	0.315	0.313	0.513 [0.412]
IAPL (Li et al., 2026)	0.920	0.713	0.715	0.582	0.500	0.500	1.000 [0.899]
InfoFD (Qin et al., 2025)	0.899	0.656	0.663	0.762	0.500	0.500	1.000 [0.899]
LaDeDa (Cavia et al., 2024)	0.822	0.608	0.612	0.549	0.500	0.500	0.993 [0.892]
LASTED (Wu et al., 2025)	0.819	0.736	0.728	0.759	0.525	0.495	0.854 [0.753]
LGrad (Tan et al., 2023)	0.745	0.638	0.635	0.581	0.500	0.500	0.993 [0.892]
MDMF (Zhang et al., 2026)	0.818	0.808	0.801	0.786	0.500	0.500	0.635 [0.534]
MIRROR (Liu et al., 2026)	0.997	0.998	0.986	0.997	0.256	0.245	0.483 [0.382]
MLEP (Yuan et al., 2025)	0.923	0.651	0.648	0.600	0.500	0.500	0.987 [0.886]
NPR (Tan et al., 2024a)	0.845	0.541	0.532	0.591	0.500	0.500	0.991 [0.890]
OmniAID (Guo et al., 2026)	0.946	0.797	0.802	0.739	0.500	0.500	0.883 [0.781]
PatchCraft (Zhong et al., 2023)	0.876	0.645	0.650	0.595	0.500	0.500	0.978 [0.877]
PGC (Zhou et al., 2026)	0.958	0.901	0.883	0.876	0.484	0.478	0.913 [0.811]
PROBE (Cao et al., 2026)	0.960	0.930	0.925	0.909	0.421	0.412	0.620 [0.518]
Raising the Bar (Cozzolino et al., 2024)	0.875	0.834	0.814	0.835	0.480	0.471	0.792 [0.690]
RealStats (Zisman & Shaham, 2026)	0.623	0.618	0.629	0.617	0.500	0.500	0.964 [0.862]
RIGID (He et al., 2024)	0.706	0.623	0.630	0.665	0.500	0.500	0.956 [0.854]
RINE (Koutlis & Papadopoulos, 2024)	0.885	0.778	0.776	0.743	0.500	0.500	0.918 [0.817]
SAFE (Li et al., 2025b)	0.881	0.589	0.583	0.579	0.500	0.500	0.992 [0.891]
SDAIE (Zhong et al., 2026)	0.915	0.717	0.716	0.579	0.500	0.500	0.993 [0.892]
SPAI (Karageorgiou et al., 2025)	0.744	0.681	0.684	0.670	0.500	0.500	0.814 [0.712]
SSP (Chen et al., 2024b)	0.783	0.592	0.594	0.575	0.500	0.500	0.958 [0.857]
Stay-Positive (Rajan & Lee, 2025)	0.875	0.786	0.736	0.745	0.500	0.500	0.928 [0.827]
Synthbuster (Bammey, 2024)	0.610	0.545	0.545	0.636	0.500	0.500	0.997 [0.896]
Synthetic Laundering (Mandelli et al., 2024)	0.756	0.719	0.714	0.696	0.500	0.500	0.824 [0.723]
UniversalFakeDetect (Ojha et al., 2023)	0.755	0.679	0.677	0.691	0.500	0.500	0.783 [0.681]
WaRPAD (Choi et al., 2025)	0.771	0.603	0.608	0.675	0.500	0.500	0.980 [0.879]
WePe (Nie et al., 2025)	0.669	0.652	0.650	0.650	0.500	0.500	0.749 [0.647]

E IMPLICATIONS

E.1 RANKING AND SHARED DECISIONS

The risk result changes how robustness claims for AI-generated image detection should be read. A condition-specific ranking asks whether one scalar orders real and generated images within a particular source, generator, and processing setting. A shared fixed threshold asks whether the same orientation and cutoff control error everywhere. These are different statistical properties. Cross-generator benchmarks, content-matched data, and consistent model selection strengthen the first type of evidence (Guillaro et al., 2025; Park & Owens, 2025; Gulrajani & Lopez-Paz, 2021); the critical shared class error measures the second directly.

This distinction matters whenever a detector is deployed with one operating point. Average area under the receiver operating characteristic curve can remain high when a small set of adverse conditions dominates worst-case risk. Per-condition recalibration can also conceal incompatible score locations because the conditions receive separate thresholds. The present result fixes both degrees of freedom and characterizes the scalar response as a shared decision statistic. Several conditions retain substantial local ordering, while the simultaneous fixed-rule criterion records their incompatibility. Future detector evaluations can report both objects: local ranking quality and simultaneous fixed-rule risk.

Across the 58 matched detectors, cellwise AUROC measures local ordering, while independently calibrated orientations and cutoffs test whether each absolute score scale supports one operating point. Their joint presentation therefore makes score portability directly measurable.

E.2 A HIERARCHY OF EVIDENCE SURVIVAL

The theory separates response survival, rule compatibility, and condition coverage. Response survival is a mean-level statement controlled by clean separation, retained processing gain, centered distortion, and non-commutation. Rule compatibility is distributional: even uniformly signed means can coexist with incompatible tails or atoms, so the exact threshold intersections remain necessary. Condition coverage asks whether represented conditions control risk elsewhere through the pseudometric $d_{\mathcal{H}}$. Each layer has its own sufficient quantities and its own empirical estimand.

This hierarchy explains why a single robustness score is inadequate. Information-contraction theory predicts that a common channel can reduce distinguishability (Raginsky, 2016), while domain-generalization theory relates transfer to domain and representation structure (Blanchard et al., 2021; Tripuraneni et al., 2020). The response-survival framework places these intuitions around a fixed scalar decision. A processing map can preserve mean separation yet shift score locations enough to destroy a shared threshold. A threshold can work on represented conditions while risk coverage remains unresolved elsewhere. Keeping the layers separate makes positive and negative results comparable.

E.3 CRITERIA FOR FUTURE RESPONSE DESIGN

The framework gives a reusable test for a candidate response q. First, estimate a clean lower separation bound for every intended source–generator pairing. Second, measure how much of that bound survives the processing maps through transport, approximation, and operator-interaction terms. Third, construct atom-aware response CDF bands and evaluate every threshold in both orientations. Finally, characterize response-space coverage if the intended condition set extends beyond the evaluated conditions. Each step can fail independently, which directs method development toward the quantity that actually limits performance.

Response alignment matters alongside magnitude: larger conditionwise mean gaps help only when score orientations and locations remain compatible. Candidate responses can therefore be selected for positive simultaneous margin and threshold-intersection slack, followed by confirmatory worst-condition evaluation. Natural-image statistics, spectral responses, and reconstruction responses provide diverse starting points (Zhang et al., 2025c; Karageorgiou et al., 2025; Ricker et al., 2024). The testable target is a response whose evidence survives processing and admits one useful fixed decision.