IDENTIFYING GENERATIVE PATH EFFECTS IN AI GENERATED IMAGE DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI generated image detectors seek statistical evidence left by modern image synthesis systems. A final image can pass through a codec (*i.e.*, an encode–decode reconstruction stage), conditioning mechanism, and a sequence of generation updates before evaluation. However, these operations may move the same score in distinct or opposing directions, so an endpoint comparison identifies only their combined effect. To this end, we construct source-matched interventions in which each image appears as a source, after a generator codec round trip, and after ordered levels of source-conditioned image-to-image generation strength, which we term doses. Potential outcomes define codec, incremental path, dose, conditional content, and cross-generator effects for any fixed detector. Complete source crossing identifies every finite linear contrast of the intervention cell means. With an arbitrary observed coordinate set, a contrast is identified exactly when its coefficient vector lies in the row space of the coordinate projection. Consequently, the source cell, every codec cell, and every generation dose cell jointly recover all codec and incremental components, requiring 21 cell means in our four generator setting. Removing codec coordinates creates one aliasing dimension per affected generator, while constructive examples establish further ambiguity for unmatched source populations and unobserved mediators. Bounded scores and independent source sampling yield simultaneous finite sample intervals; the same estimators are exact for a frozen roster. Finally, we instantiate the theory with 58 public detectors, four versioned generators, four doses, and 2,048 sources. All 3,132 primary contrasts are identified, revealing path sensitivity, generator heterogeneity, and causal component cancellation. In addition, a text to image comparison examines the same detector profiles on complete synthesis routes.

1 INTRODUCTION

AI generated image detection asks whether an image carries evidence of modern synthesis. The question becomes increasingly important as generators produce diverse, photorealistic images that enter ordinary media collections. Public detectors approach this evidence through markedly different response functions: some use frozen vision representations (Ojha et al., 2023), some use diffusion inversion (Wang et al., 2023), and others use autoencoder reconstruction (Ricker et al., 2024), spectral structure (Karageorgiou et al., 2025), natural image distributions (Zhang et al., 2025b), or real image reference statistics (Zisman & Shaham, 2026). Each method ultimately compresses the observed image into a score that supports a detection decision.

However, the image reaching a detector can have a long generative history. A source may pass through an encoder and decoder and then undergo conditioned generation at one of several strengths. Each operation can alter the same detector score. A conventional difference between natural and generated images merges these changes. That means, if codec passage raises the score and subsequent generation lowers it by a comparable amount, the final comparison appears small even though the detector responds strongly to both operations. A large final difference may likewise arise almost entirely from codec passage. The measured gap thus leaves its component causes ambiguous. Fig. 1 illustrates this ambiguity: a modest source-to-final score change can conceal large, opposing effects from codec passage and subsequent generation. This ambiguity motivates our central question: *which component of a versioned generative path changes a fixed detector score, and which experimental cells make that effect identifiable?*

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.

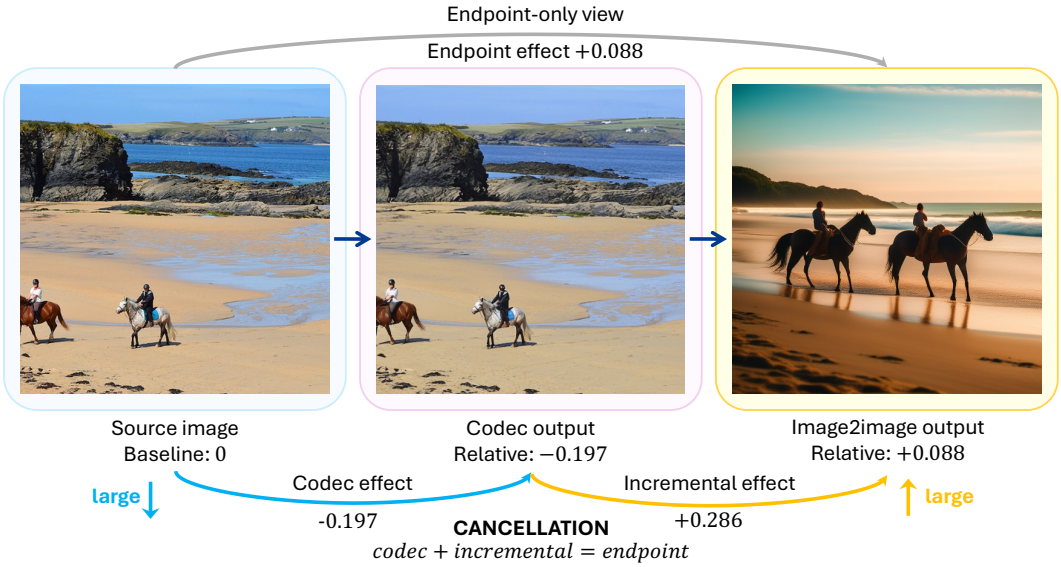

Figure 1: An endpoint comparison conceals opposing changes along a generative path. The three images illustrate the source, codec, and generated stages; and the reported effects are averages over all 2,048 sources.

Recent benchmarks broaden generator coverage and expose content, format, and dataset shortcuts (Guillaro et al., 2025; Yan et al., 2025b; Park & Owens, 2025; Yang et al., 2025). A comparison between natural photographs and text to image outputs provides a direct evaluation of complete synthesis routes. However, locating the source of a score change additionally requires intermediate observations in which content and path components vary in a controlled way. Source-conditioned image to image generation provides this structure: the same image and prompt can enter a source cell, a generator codec round trip, and a sequence of ordered generation doses. The resulting comparisons identify where detector evidence changes along the path. A separate complete route comparison then connects these source-matched response profiles to text to image synthesis.

We formalize this construction by treating every trained detector as a fixed measurement function inside a source-matched causal experiment. Each source image and its fixed prompt enter a source control, one codec round trip for each generator family, and four ordered image to image doses. Two scheduled seeds realize every dose. Comparisons occur within the same source, so source content remains paired across all interventions. Detector scoring is blind to treatment labels, and the labels enter only when the complete score table joins the frozen manifest. This design turns detector evaluation from a comparison of image collections into a comparison of controlled potential outcomes (Dasgupta et al., 2015).

The resulting decomposition has a simple interpretation. For detector j and generator g, the codec effect τ_{jg}^{dec} measures the score change after the registered codec round trip. The source relative dose effect $m_{jg}(d)$ measures the change after the complete source-conditioned path at dose d. Their difference, $\eta_{jg}(d) = m_{jg}(d) - \tau_{jg}^{\mathrm{dec}}$, isolates the incremental path that follows codec passage. Adjacent dose effects describe how the response develops within a generator family, and differences between codec effects describe cross-generator heterogeneity. Opposing values of τ^{dec} and η reveal cancellation: a modest end to end score change can conceal two large responses with opposite directions. These objects connect a practical question about detector behavior to explicit contrasts between controlled image interventions.

Our theory states exactly when these quantities are recoverable. Complete source crossing identifies every finite linear contrast of the intervention cell means. For four generators and four doses, 21 cell means form the smallest coordinate design that recovers every codec and incremental component. Hiding one codec cell preserves the total source relative dose effect while allowing the codec and incremental components to shift in opposite directions along an aliasing dimension. An unmatched

observational comparison introduces a second ambiguity because intervention and source content vary together. Constructive witnesses make both ambiguity directions explicit. The same potential outcome framework also supports content conditioned effects and simultaneous finite sample intervals that preserve dependence among doses, seeds, paths, and detectors sharing a source. Identification therefore becomes a property of the intervention design and target contrast, with a precise scope for the final score difference.

We instantiate the design with 2,048 MSCOCO source images (Lin et al., 2014), four versioned generator families, four ordered doses, and 58 publicly available AI generated image detectors. The crossing contains 75,776 images and 4,395,008 detector scores. It identifies all 3,132 predeclared contrasts, comprising codec, source relative dose, incremental path, adjacent dose, and cross-generator effects. Every point estimate is an exact mean for the frozen source roster. Simultaneous intervals provide the corresponding analysis under the stated independent source sampling model. Treatment blind scoring and source level averaging preserve the experimental pairing from image construction through statistical analysis. A secondary text to image analysis uses the same generator families and detector panel to examine whether the resulting response profiles extend from controlled source-conditioned paths to complete synthesis routes.

The results show that the generative path matters at the level of individual detector scores. Forty of 58 detectors have at least one codec interval that excludes zero, 40 have at least one incremental path interval that excludes zero, and 35 have codec effects that vary across generators. Among 928 matched path decompositions, 261 contain codec and incremental components with opposite signs. Both directions are simultaneously resolved in 24 cases, and 50 cases satisfy a strict cancellation criterion. For FreqDetect (Frank et al., 2020) under Kandinsky 2.2 (Kandinsky Community, 2023) at dose 0.8, a codec effect of -0.197 and an incremental effect of 0.286 combine into a source relative effect of only 0.088. Detector profiles also separate qualitatively: FerretNet (Liang et al., 2025) responds strongly to codec passage with little resolved incremental response, while FatFormer (Liu et al., 2024) shows the reverse pattern in several generation cells. These findings explain why the same final score change can represent distinct generative causes. They also show how a source-matched experiment reveals detector sensitivities that remain hidden in an aggregate real versus generated comparison. This paper makes three contributions:

- We formulate fixed detector scores as potential outcomes and define identifiable codec, dose, incremental path, adjacent dose, and cross-generator effects.
- We establish complete crossing identification, a minimal 21 cell coordinate design, constructive nonidentification for missing and observational comparisons, and simultaneous source sampling intervals.
- We measure 58 publicly available detectors across 2,048 sources and four generators, identify 3,132 primary contrasts, and assess complete text to image routes in a secondary validation.

2 Related Work

2.1 Generalizable AI generated image detection

AI generated image detectors learn or construct evidence from artifacts, representations, reconstruction behavior, and natural image statistics. Frozen CLIP features transfer across generators (Ojha et al., 2023), while diffusion based approaches use inversion trajectories (Wang et al., 2023). Autoencoder reconstruction exposes discrepancies between real and generated images (Ricker et al., 2024). Other recent methods model spectral responses (Karageorgiou et al., 2025), deviations from natural image distributions (Zhang et al., 2025b), or reference statistics learned from real images alone (Zisman & Shaham, 2026). We use these and other public methods as fixed measurement functions. Specifically, their scores become outcomes of controlled generative interventions, allowing the same causal question to be asked across a diverse detector panel.

2.2 Data design and detector evaluation

Evaluation quality depends strongly on how real and generated collections are assembled. Content alignment and augmentation reduce shortcuts caused by semantic composition and file format (Guillaro et al., 2025). Evaluation on recent high quality images reveals behavior hidden by conventional benchmark splits (Yan et al., 2025b). Broader generator collections improve empirical coverage, and

studies of architectural diversity show that generator count alone gives an incomplete description of transfer (Park & Owens, 2025; Yang et al., 2025). Large public benchmarks consequently provide increasingly realistic estimates of detection performance. However, their usual comparison unit is still a real collection paired with a generated collection. Specifically, that unit combines all operations in the generative path and can also combine differences in source content. In contrast, our design changes the comparison unit to one source observed under a complete crossing of source, codec, and ordered generation interventions.

2.3 Causal measurement of model behavior

Potential outcomes formalize treatment effects in factorial experiments and make the required assignment structure explicit (Dasgupta et al., 2015). Counterfactual image interventions support the measurement of classifier bias (Dash et al., 2022), and causal audits connect controlled perturbations to model robustness (Drenkow et al., 2026). Causal representation research studies the conditions under which learned variables support intervention level interpretation (Wang & Jordan, 2024; Ahuja et al., 2023). These lines establish that causal conclusions depend jointly on the target contrast and the available interventions. We apply this principle to generative paths preceding detector scores. Specifically, complete source crossing identifies codec and subsequent path effects for fixed detectors, while a row-space criterion and constructive witnesses characterize the ambiguity caused by missing codec cells or unmatched source populations.

3 Causal Effects of Generative Paths

An ordinary real versus generated comparison combines several generative operations into one detector score difference. In contrast, potential outcomes represent the score that a fixed detector produces when each operation is applied separately (Dasgupta et al., 2015). Comparing these scores answers two questions: *which operation changes the detector response, and by how much?*

We instantiate this idea along a two stage generative path. A source image first passes through the codec of a generative model and then undergoes source-conditioned generation. The source image, codec output, and final image form three experimental cells. Score contrasts among these cells isolate the codec effect, the subsequent generation effect, and their combined effect. Repeating the construction for every source, generator, and dose creates a vector of potential detector scores whose linear contrasts have direct causal meanings.

The below theory proceeds in four steps. First, it defines intervention cells and their source averaged detector outcomes. Second, it expresses codec, total, incremental, dose, and generator effects as contrasts of the corresponding cell means. Third, it establishes which contrasts the complete crossed design identifies and characterizes the ambiguity created by missing cells or unmatched source populations. Fourth, it gives simultaneous uncertainty statements and continuous coordinates that summarize the resulting detector response geometry. Each step has a direct experimental counterpart in Section 4 and the confirmatory results.

3.1 Source units, interventions, and outcomes

A source unit is $Z = (X, P, C)$, where X is an image, P is its fixed prompt, and C contains preintervention content descriptors. Let $\mathcal{G}$ be a finite nonempty set of generator families. Generator g has $K_g \geq 1$ ordered image to image doses $\mathcal{D}_g = \{d_{g1} < \cdots < d_{gK_g}\}$. The intervention cells are

$$\mathcal{T} = \{t_0\} \cup \{t_g^{\mathrm{dec}} : g \in \mathcal{G}\} \cup \{t_{gk}^{\mathrm{i2i}} : g \in \mathcal{G}, k = 1, \ldots, K_g\}, \tag{1}$$

where $t_0 = (\varnothing, \mathrm{source}, 0)$, $t_g^{\mathrm{dec}} = (g, \mathrm{decoder}, 0)$, and $t_{gk}^{\mathrm{i2i}} = (g, \mathrm{img2img}, d_{gk})$. The source cell executes no generator, so its image is invariant to a nominal generator label:

$$I_{(g,\mathrm{source},0)}(z; \varnothing) = I_{(\varnothing,\mathrm{source},0)}(z; \varnothing). \tag{2}$$

For each cell t, let $\boldsymbol{\xi}_t(z) = (\xi_{t1}(z), \ldots, \xi_{tR_t}(z))$ denote a replicate schedule fixed from preintervention source identity before detector scores are available. Let $I_t(z; \xi_{tr}(z))$ be the resulting image and s_j the frozen scalar score of detector j. The scheduled potential score, replicate mean, and population cell mean are

$$Y_{jr}(z,t) = s_j(I_t(z;\xi_{tr}(z)))\,, \qquad \overline{Y}_j(z,t) = \frac{1}{R_t}\sum_{r=1}^{R_t} Y_{jr}(z,t), \qquad \mu_j(t) = \mathbb{E}_Z\left[\overline{Y}_j(Z,t)\right]. \quad (3)$$

Source and decoder cells use $R_t = 1$ and $\xi_{t1}(z) = \varnothing$. Image to image cells use the two integer seeds in their frozen schedule.

The expectation over Z refers either to a stated superpopulation or to the uniform distribution on a frozen finite-roster. The latter interpretation limits the estimand to that roster.

In the experiment, each treatment blind score is associated with one source and one intervention cell. Source and decoder cells have one realization. An image to image cell has $R = 2$ scheduled replicates. The observed cell outcome first averages those replicates within source:

$$\widetilde{Y}_{ij}(t) = \frac{1}{R_t}\sum_{r=1}^{R_t} Y_{ijr}^{\mathrm{obs}}(t), \qquad R_t = \begin{cases} 1, & t = t_0 \text{ or } t = t_g^{\mathrm{dec}}, \\ R, & t = t_{gk}^{\mathrm{i2i}}. \end{cases} \quad (4)$$

All later contrasts are formed from $\widetilde{Y}_{ij}(t)$ inside source unit i. Under consistency, this quantity equals $\overline{Y}_j(z_i,t)$ exactly. The target is the arithmetic mean over the two frozen replicates; a full generator-randomness target requires a separate random-seed design. This order preserves dependence among doses, paths, seeds, and detectors that share the same source.

3.2 Path and dose estimands

For coefficients $c = (c_t : t \in \mathcal{T})$ with $\sum_t c_t = 0$, define the general linear estimand

$$\theta_{j,c} = \sum_{t\in\mathcal{T}} c_t \mu_j(t). \quad (5)$$

The manifest label "decoder" denotes the complete deterministic codec round trip defined above. The codec round trip effect, source relative dose response, and incremental image to image contrast are

$$\tau_{jg}^{\mathrm{dec}} = \mu_j(t_g^{\mathrm{dec}}) - \mu_j(t_0), \qquad m_{jg}(d_{gk}) = \mu_j(t_{gk}^{\mathrm{i2i}}) - \mu_j(t_0). \quad (6)$$

$$\eta_{jg}(d_{gk}) = \mu_j(t_{gk}^{\mathrm{i2i}}) - \mu_j(t_g^{\mathrm{dec}}) = m_{jg}(d_{gk}) - \tau_{jg}^{\mathrm{dec}}. \quad (7)$$

The term $\eta_{jg}(d)$ remains an incremental path contrast. The empirical image to image cells add latent preparation, scheduled randomness, prompt conditioning, guidance, and denoising; Kandinsky also adds its fixed prior. A denoising-specific effect requires a different modular intervention.

A simple source level comparison gives the interpretation. Consider a detector whose score changes when the source passes through a generator codec and then changes again when the decoded image enters the remaining source-conditioned generation path. The source to final comparison measures $m_{jg}(d)$. The source to codec comparison measures τ_{jg}^{dec}, and the codec to final comparison measures $\eta_{jg}(d)$. Their identity $m_{jg}(d) = \tau_{jg}^{\mathrm{dec}} + \eta_{jg}(d)$ shows why one total change can conceal two substantial components. Components with the same sign reinforce each other, while components with opposite signs produce cancellation. The matched construction makes all three comparisons for the same source unit, so stable source content enters every term in the same way.

The remaining contrasts answer complementary questions. Adjacent dose effects describe how detector response changes as the registered generation dose increases within one family. Codec heterogeneity compares how different generator implementations reconstruct the same sources before their later generation steps occur. Conditional effects repeat a path contrast within a preintervention content stratum. Together, these estimands distinguish the location of a detector response along the generative path, its variation with dose, its dependence on generator identity, and its dependence on source content.

Adjacent doses reveal the shape of the response without a monotonicity assumption:

$$\Delta_{jg,k} = \mu_j(t_{gk}^{\mathrm{i2i}}) - \mu_j(t_{g,k-1}^{\mathrm{i2i}}), \quad k = 2,\ldots,K_g, \qquad \Gamma_{j;g,g'}^{\mathrm{dec}} = \tau_{jg}^{\mathrm{dec}} - \tau_{jg'}^{\mathrm{dec}}. \quad (8)$$

For two generator families, $\Gamma_{j;g,g'}^{\mathrm{dec}}$ measures codec round trip heterogeneity.

Cross-family dose contrasts require a separately justified common dose scale. The primary analysis consequently uses Eq. 8 for cross-family comparisons and retains dose curves within family.

Content C may modify any linear path effect. For coefficients $\ell = (\ell_t : t \in \mathcal{T})$ satisfying $\sum_t \ell_t = 0$ and a predeclared discrete stratum $C = c_0$ with positive probability under the stated target, the conditional estimand is

$$\theta_{j,\ell|c_0} = \mathbb{E}\left[\sum_{t\in\mathcal{T}} \ell_t \overline{Y}_j(Z,t) \,\middle|\, C = c_0\right]. \tag{9}$$

Source matching permits heterogeneous source content because every comparison occurs within Z. The marginal effect still depends on the roster's content mixture, and Eq. 9 makes this dependence observable.

3.3 IDENTIFICATION

The identification conditions fall into three conceptual groups. The first group gives every intervention a stable meaning: implementations, replicate schedules, detector mappings, and observed potential outcomes remain consistent. The second group supplies the comparison structure: each source has every required cell, source units satisfy noninterference, and replicate seeds are assigned before detector outcomes enter the analysis. The third group defines the population target and its regularity: sources represent either an independent sample from a declared population or the complete finite-roster, and the relevant score moments exist. This grouping separates the causal design requirements from the population interpretation and the measurement requirements.

The result uses eight conditions. Each cell has a fixed implementation and replicate schedule (A1). Observed outcomes satisfy consistency (A2). Source-level noninterference holds (A3). Sources are independent draws from the target distribution, or the fixed roster itself defines the target (A4). Every required cell appears for every source (A5). The replicate schedule is fixed before detector outcomes (A6). Detector code and score mapping remain fixed (A7). Cell outcomes have finite first moments (A8). Appendix E.1 gives their formal roles.

Theorem 1 (Identification by complete source crossing)**.** *Under A1 through A8, every finite linear cell-mean contrast in Eq. 5 is identified by the joint distribution of the crossed source-level outcome vector. The estimator*

$$\widehat{\theta}_{j,c} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t\in\mathcal{T}} c_t \widetilde{Y}_{ij}(t) \tag{10}$$

is unbiased under superpopulation sampling and converges almost surely to $\theta_{j,c}$. For a fixed roster it equals the roster-average scheduled effect exactly.

Theorem 1 applies directly to Eqs. 6 through 8. Its experimental content is concrete: a valid manifest supplies every coefficient's cell, treatment-blind scoring supplies the outcomes, and the source identifier defines the summation unit in Eq. 10. The proof appears in Appendix E.4.

3.4 WHAT INCOMPLETE DATA LEAVE UNRESOLVED

The codec cell supplies the reference point that locates the incremental effect. If an experiment records only a source score and a final image to image score, their difference identifies the combined effect $m_{jg}(d)$. Many pairs of codec and incremental effects yield that same combined value: increasing the codec component by an amount and decreasing the incremental component by the same amount preserves the observed total. Observing the codec score fixes this otherwise free direction. This geometric ambiguity motivates the general identification statement below and the cell minimality result that follows it.

Theorem 2 (Constructive nonidentification)**.** *Three limits hold for an unrestricted potential-outcome model.*

(a) Without exchangeability and overlap, an observational real or generated association admits two causal models with the same observed distribution and different average generative-path effects.

(b) Let P_Ω select the observed coordinates of the cell-mean vector μ. Suppose the mean model is variation independent, $\mathcal{M} = \prod_{t\in\mathcal{T}} \mathcal{I}_t$, where every $\mathcal{I}_t$ is a nondegenerate interval. A linear contrast $c^\top \mu$ is identified over $\mathcal{M}$ exactly when

$$c \in \mathrm{row}(P_\Omega) = \ker(P_\Omega)^\perp. \tag{11}$$

(c) Observed treatment cells and scalar detector scores alone leave the decomposition of a score effect into unobserved semantic and artifact mediators unidentified.

The three parts separate population comparability, observed coordinate coverage, and mediator attribution. Their corresponding constructions appear in Appendix E.5.

The missing-decoder case illustrates part (b). Source and image to image cells identify $m_{jg}(d)$, while the components satisfy only

$$m_{jg}(d) = \tau_{jg}^{\mathrm{dec}} + \eta_{jg}(d). \tag{12}$$

The prespecified incomplete view hides every FLUX.1 codec coordinate while retaining the source and image to image cells. It preserves $m(d)$ while the two components vary along the aliasing direction in Eq. 12.

Corollary 1 (Cell minimality and decoder aliasing)**.** *For an unrestricted cell mean model, the joint codec and incremental effect vector is identified exactly when the source, every codec, and every image to image cell are observed. These* $1 + |\mathcal{G}| + \sum_g K_g$ *cell means form a minimal coordinate design. If codec means are missing for* q *families, the compatible component vectors contain a* q *dimensional affine patch while every source relative effect remains fixed.*

The confirmatory design observes all 21 required cell means per source. Please refer to Appendix E.2 for the full algebraic statement and to Appendix E.5 for the constructions and proof.

3.5 Simultaneous finite-sample uncertainty

For bounded detector scores and independent source sampling, one simultaneous event covers every predeclared contrast while preserving dependence among all measurements of the same source. For the finite-roster target, the same estimators are exact. Please refer to Appendix E.6 for the formal proposition, radius, and proof (Romano & Wolf, 2005).

Please refer to Appendix E for the complete assumptions and proofs. The signed contrast vector remains the inferential target; descriptive coordinates D_j, V_j, H_j summarize its geometry. Please refer to Appendix A for their definitions and domains.

4 Experimental Design

We instantiate the crossed design with 2,048 MSCOCO source images (Lin et al., 2014). Each source supplies one fixed prompt and enters four versioned generator families: FLUX.1 dev, Qwen Image, Kandinsky 2.2, and Kolors (Black Forest Labs, 2024; Wu et al., 2025a; Kandinsky Community, 2023; Kolors Team, 2024). For every family, the observed cells contain the source image, the registered codec round trip, and source-conditioned image to image generation at doses $0.2, 0.4, 0.6, 0.8$. Two source specific seeds realize each dose and remain paired across the dose sequence.

The crossing produces 75,776 images and 4,395,008 scores from 58 publicly available detectors. Each paper level method contributes one canonical public inference route, frozen preprocessing, and a scalar outcome oriented toward generated image evidence. Treatment labels remain hidden during scoring. The analysis averages the paired seeds within source and forms all contrasts within the same source unit. This construction yields 3,132 predeclared primary contrasts spanning codec, source relative, incremental, adjacent dose, and cross-generator effects. Point estimates are exact for the frozen roster; simultaneous intervals use the source sampling interpretation in Proposition 1.

Please refer to Appendix B for source selection, captioning, model revisions, seed construction, detector score mappings, reference banks, the complete contrast family, and the content conditioned analyses. Appendix F gives the primary reference and public code or model repository for every evaluated detector.

Table 1: Main empirical findings. Each denominator is the complete declared family for that statement.

Finding	Result	Finding	Result
Identified primary contrasts	3,132 / 3,132	Opposing path components	261 / 928
Resolved codec response	40 / 58	Resolved opposing components	24 / 928
Resolved incremental response	40 / 58	Strict cancellation	50 / 928
Cross-generator codec heterogeneity	35 / 58	Resolved text-to-image route contrasts	32 / 232

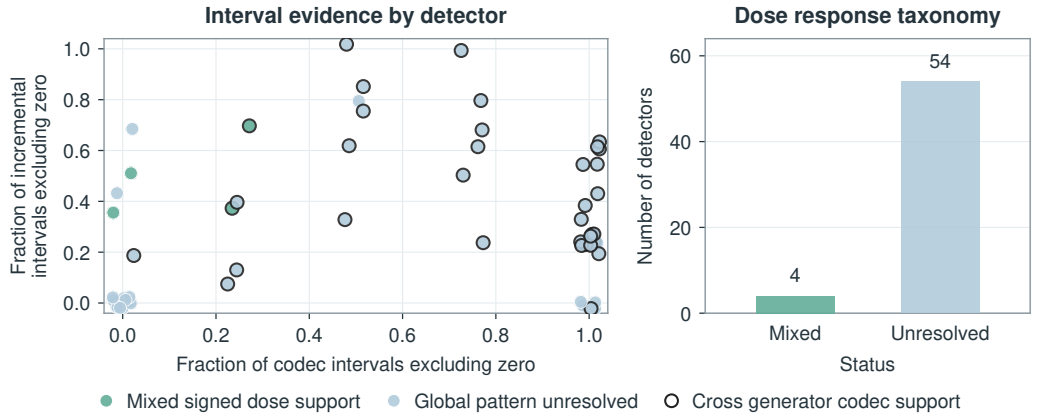

Figure 2: Detector-level response structure under the complete crossed design. Interval classifications use A3 and A4(i); D, V, H are descriptive finite-roster coordinates.

5 Results

Table 1 provides an overview of the main empirical findings; the observations below unpack their path-level structure.

Observation 1: Detector sensitivity is widespread across both observed stages of the generative path, while local dose changes are much less consistently resolved. The complete crossing identifies all 3,132 primary contrasts. Under the stated source-sampling model, simultaneous intervals exclude zero for at least one codec effect in 40 of 58 detectors, at least one incremental path effect in 40, and at least one cross-generator codec contrast in 35. Across individual contrasts, intervals exclude zero for 127 of 232 codec effects, 606 of 928 source-relative effects, 298 of 928 incremental effects, 53 of 696 adjacent-dose effects, and 122 of 348 cross-generator codec contrasts. Thus complete paths frequently move detector scores, whereas resolved local changes between adjacent doses are comparatively uncommon. Fig. 2 summarizes this detector-level response structure.

Observation 2: Endpoint effects can conceal large and opposing responses at different stages of the same generative path. Among 928 detector, generator, and dose triples, 261 contain codec and incremental effects with opposite signs. Both directions are simultaneously resolved in 24 cases, and 50 cases satisfy the strict cancellation criterion $|m| < \min\{|\tau^{\mathrm{dec}}|, |\eta|\}$. For FreqDetect under Kandinsky 2.2 at dose 0.8, a codec effect of -0.197 and an incremental effect of 0.286 combine into a source-relative effect of only 0.088. The modest endpoint response therefore conceals two substantial effects in opposite directions. MIRROR illustrates the reinforcing case: all 16 incremental intervals lie above zero, while three of four codec intervals lie above zero.

Observation 3: Different detectors respond to different locations along the generative path. FerretNet has four resolved codec effects between 0.966 and 0.983, while all 16 of its incremental intervals include zero. FatFormer exhibits the complementary pattern: its four codec intervals include zero and eight of its 16 incremental intervals lie above zero. AlignedForensics shows strong generator dependence: its codec effect changes from 0.012 under FLUX.1 to 0.982 under Kandinsky 2.2, producing a cross-generator contrast of -0.971. MIRROR responds at both observed stages, with three of four codec effects and all 16 incremental effects resolved. These representative profiles are summarized in Table 2. Across the complete panel, four methods show resolved adjacent-dose changes in both directions, while the remaining 54 have an unresolved global dose pattern.

Table 2: Representative detector response profiles. Support entries give the number of simultaneous intervals excluding zero over the declared family.

Detector	Codec	Incremental	Cross-generator	Profile
FerretNet	4/4	0/16	0/6	Codec concentrated
FatFormer	0/4	8/16	0/6	Incremental concentrated
AlignedForensics	3/4	11/16	5/6	Generator dependent
MIRROR	3/4	16/16	5/6	Reinforcing stages

Observation 4: Intermediate codec cells and source matching provide identifying information that endpoint comparisons alone cannot recover. Two candidate hidden codec means produce distinct codec and incremental components while reconstructing every observed source-relative effect to within 1.11×10^{-16}. Thus the same endpoint effects admit different path decompositions when the intermediate codec coordinate is unobserved. Separately, reversing the caption-content allocation changes the sign of the matched FLUX.1 dose 0.8 effect for four detectors in one unmatched comparison and three in the other. The two constructions expose distinct sources of ambiguity: missing intermediate cells leave the path decomposition unresolved, while unmatched source populations can alter the apparent effect itself.

Observation 5: Aggregate response structure remains stable across the tested content split and across complete synthesis routes, despite detector-specific deviations. For the predeclared caption split, the median FLUX.1 dose 0.8 effect is 0.156 among 865 captions that mention a person and 0.167 among 1,183 captions without a person term. Across the 58 detectors, the median stratum difference is 0.0003, with an interquartile range of 0.0115. Across the four generators, detector-profile correlations between source-conditioned and text-to-image routes range from 0.977 to 1.000, while 32 of 232 route contrasts have simultaneous intervals excluding zero. The aggregate similarity therefore coexists with detector-level variation in both analyses. Appendix I reports the family-level route results, and Appendix J gives all detector-level values.

Additional directional structure: resolved within-path effects are predominantly positive, whereas cross-generator codec differences are more balanced in sign. Among codec effects, 119 intervals lie above zero and eight lie below; among incremental effects, 279 lie above and 19 lie below. Source-relative effects contain 578 positive and 28 negative resolved intervals, whereas adjacent-dose effects contain only 43 positive and ten negative intervals. Cross-generator codec contrasts are more balanced, with 58 positive and 64 negative resolved intervals. The largest codec effect is 0.987 for PROBE under Kolors, and the largest incremental effect is 0.753 for FatFormer under Kolors at dose 0.6. Across all 928 decompositions, the maximum numerical residual in $m = \tau^{\mathrm{dec}} + \eta$ is 3.43×10^{-15}, confirming the decomposition identity to numerical precision.

Appendix C reports the complete estimand partitions, effect directions, detector profiles, observational comparison, and content-conditioned results. Complete assumptions and proofs appear in Appendix E. Appendix G gives every primary detector contrast and its interval status, Appendix H reports all predeclared content-stratum effects, and Appendix D discusses consequences for benchmark and detector design.

6 Conclusion

An AI generated image detector can respond to several operations along one generative path. Source-matched interventions separate these operations into codec round trip, source relative, incremental, adjacent dose, and cross-generator effects on a fixed detector score. The potential outcome framework establishes when the complete crossed design identifies these effects, why its cell coordinates are minimal, and how a missing codec cell creates an affine ambiguity between codec and subsequent generation responses. Across 58 public detectors, both path stages commonly change scores, the effects vary across generator implementations, and opposing components frequently combine within one end to end change. The text to image extension shows that route choice also produces substantial detector specific responses. These findings connect the formal identification results to practical detector evaluation: measuring intermediate path cells turns an aggregate score difference into interpretable causal contrasts and reveals response structure that an endpoint comparison conceals.

AI USE STATEMENT

We use generative AI tools in supporting roles for aspects of research execution and manuscript preparation, including language editing, manuscript restructuring, and code generation and debugging. All mathematical formulations and equations are developed and drafted manually by the authors, and the associated analyses and scientific interpretations are conducted manually by the authors. GPT-5.6 Sol is used as a secondary check of the mathematical expressions for correctness and internal consistency; the authors independently review and determine the final formulations. The generative image models and vision-language captioning model described in the experimental protocol are experimental systems/interventions in this study. No generative AI assistant was used to make final scientific decisions or independently interpret experimental results. All AI-assisted code is reviewed, tested, and validated against the intended computations and experimental outputs. We likewise review and revise all AI-assisted manuscript text and check reported numerical results against the underlying experimental outputs. The authors take full responsibility for the final content of this work, including its text, claims, mathematical formulations, code, analyses, and other artifacts produced with the aid of generative AI.

REFERENCES

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 372–407. PMLR, 2023.

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.

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.

Black Forest Labs. FLUX.1-dev model card, 2024. Public model release.

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. *arXiv preprint arXiv:2605.24906*, 2026.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real world. *arXiv preprint arXiv:2406.09398*, 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*, pp. 12830–12839, 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*, pp. 4356–4366, 2024.

Tirthankar Dasgupta, Natesh S. Pillai, and Donald B. Rubin. Causal inference from 2^K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 77(4):727–753, 2015. doi: 10.1111/rssb.12085.

Saloni Dash, Vineeth N. Balasubramanian, and Amit Sharma. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 915–924, January 2022.

Nathan Drenkow, William Paul, Chris Ribaudo, and Mathias Unberath. Causality-driven audits of model robustness. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5956–5966, March 2026.

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.

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.

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.

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.

Kandinsky Community. Kandinsky 2.2 decoder and image prior model cards, 2023. Public model release.

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.

Kolors Team. Kolors: Effective training of diffusion model for photorealistic text-to-image synthesis. *arXiv preprint*, 2024.

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.

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, 2025.

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*, pp. 21263–21272, 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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.

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*, pp. 10770–10780, 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.

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.

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.

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*, pp. 43008–43017, 2026.

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.

Joseph P. Romano and Michael Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108, 2005. doi: 10.1198/016214504000000539.

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*, pp. 3418–3432, 2023. doi: 10.1145/3576915.3616588.

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.

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.

Yixin Wang and Michael I. Jordan. Desiderata for representation learning: A causal perspective. *Journal of Machine Learning Research*, 25(275):1–65, 2024.

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.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

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

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4818–4829, 2024.

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 Generalized 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.

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, 2025a.

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, 2025b. 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, 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.

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 Detector Response Coordinates and Empirical Roles

The signed contrasts remain the inferential targets. Three continuous coordinates summarize their geometry:

$$D_j^2 = |\mathcal{G}|^{-1} \sum_{g \in \mathcal{G}} \left(\tau_{jg}^{\mathrm{dec}}\right)^2, \quad V_j^2 = \left[\sum_g (K_g - 1)\right]^{-1} \sum_{g \in \mathcal{G}} \sum_{k=2}^{K_g} \Delta_{jg,k}^2, \quad H_j^2 = |\mathcal{P}|^{-1} \sum_{(g,g') \in \mathcal{P}} \left(\Gamma_{j;g,g'}^{\mathrm{dec}}\right)^2. \tag{13}$$

Here $\mathcal{P}$ contains all six unordered pairs of the four registered generator families. D_j measures codec round trip response magnitude, V_j measures within family dose variation, and H_j measures cross family codec heterogeneity. These are descriptive plug in summaries. The complete signed vector carries the frozen roster conclusions, and the A3/A4(i) conditional intervals provide the corresponding sensitivity analysis.

The three coordinates compress different axes of the complete signed response vector. A large D_j indicates that codec passage alone moves detector scores across the registered generators. A large V_j

Table 3: Theoretical objects and their empirical roles.

Object	Experimental realization
$Y_{jr}(z,t)$	Detector score measured under a source, codec, or image to image treatment, with scheduled replicate averaging.
τ_{jg}^{dec}	Within source codec round trip score minus source score.
$m_{jg}(d), \eta_{jg}(d)$	Replicate mean image to image score minus the source or codec score.
$\Delta_{jg,k}$	Within source difference between consecutive dose means.
$\Gamma_{j;g,g'}^{\mathrm{dec}}$	Difference between two within source codec round trip effects.
$\theta_{j,\ell\mid c_0}$	Stratified extension for a preintervention content descriptor.
Theorem 2(a)	Two unmatched caption content mixtures compared with the source-matched effect.
Theorem 2(c)	Boundary on mechanism attribution from scalar scores alone.
Proposition 1	A3/A4(i) conditional intervals over each declared contrast family.
D_j, V_j, H_j	Root mean squares of the corresponding signed contrast estimates.
Corollary 1	Complete 21 cell mean vector and a prespecified incomplete view with one codec coordinate hidden.

indicates substantial variation between consecutive dose cells, including curves whose local changes have different signs. A large H_j indicates that codec response depends strongly on generator family. The confirmatory response summary places these coordinates beside the fractions of resolved codec and incremental effects, so the descriptive geometry and simultaneous interval conclusions remain visible together. These coordinates are defined when $|\mathcal{G}| > 0$, $\sum_g (K_g - 1) > 0$, and $|\mathcal{P}| > 0$, respectively. A coordinate is unavailable when its component set is empty. Table 3 distinguishes direct measurements, constructive analyses, assumption conditional intervals, and theoretical boundaries.

B Experimental Design

The experiment asks whether a detector's response to a complete generative path can be separated into effects of the codec round trip and effects of the subsequent source-conditioned generation. Answering this question requires more than a collection of real and generated images. The same source must enter every intervention, the codec must appear as its own cell, generation dose must vary within a fixed family, and the detector must remain unchanged across those cells. Our design builds these requirements into the image construction before detector scores become available.

B.1 Sources and prompts

The source image is the experimental unit. Holding it fixed across treatments removes source content from each within source contrast and makes the path operation the varying factor. The confirmatory source population is a frozen set of 2,048 images drawn by a deterministic hash ordering from 10,000 verified MSCOCO test images (Lin et al., 2014). A separate ordered set contains 256 reserve images. Every file is decoded, rehashed, and compared against both the 256 pilot sources and a separately frozen exclusion set of 14,534 source hashes. All intersections are empty. Because selection uses image bytes alone, the roster is fixed before every generative intervention and detector outcome.

The prompt is also held fixed within a source. MSCOCO test images have no released content captions, so we generate one source caption with Florence 2 large ft (Xiao et al., 2024) at a fixed model revision. The captioner sees only the source image and uses its short caption task with deterministic three beam decoding. Two executions on the same 16 images produce identical output bytes. All 2,304 analysis and reserve images then satisfy the frozen language and formatting contract. The original 2,048 analysis ranks remain intact, and no reserve replacement occurs. Each accepted caption becomes the source's prompt verbatim in every generator, dose, and seed cell. Consequently, a dose contrast changes the registered image intervention while preserving its source image and linguistic condition.

B.2 Crossed generative interventions

The treatment crossing is chosen to expose the components in Eqs. 6 through 8. The four generator families are FLUX.1 dev, Qwen Image, Kandinsky 2.2, and Kolors (Black Forest Labs, 2024; Wu

Table 4: Confirmatory crossed intervention design.

Cell	Per source	Images	Identified quantity
Source image	1	2,048	Reference
Codec round trip	4×1	8,192	τ^{dec}
Image to image doses	$4 \times 4 \times 2$	65,536	m, η, Δ
All image cells	37	75,776	Complete crossing
Detector outcomes	37×58	4,395,008	3,132 contrasts

et al., 2025a; Kandinsky Community, 2023; Kolors Team, 2024). Each family contributes one codec passage and four source-conditioned generation doses at native strengths $0.2, 0.4, 0.6, 0.8$. The source cell anchors every effect. The codec cell measures the image change produced before source-conditioned generation, and the four dose cells measure complete paths of increasing native strength. Together, these cells allow the complete source relative response to be decomposed into a codec round trip effect and an incremental path contrast.

Table 4 connects each image class to the effect it identifies and gives the complete image and score counts.

The codec cell applies the registered image processor, encoder, deterministic latent rule, decoder, and postprocessor from the loaded image to image pipeline. The Qwen Image, Kandinsky, and Kolors implementations expose the same codec objects in the codec and generation cells, and repeated executions reproduce every dose output. FLUX.1 uses the same registered VAE in both cell classes, and an independent batch 32 execution reproduces every dose 0.4 output. This shared implementation makes the codec cell a measured component of the corresponding complete path instead of a separately chosen surrogate.

Every dose requests 20 decoder steps. The four strengths execute 4, 8, 12, and 16 decoder denoising updates. Kandinsky additionally executes a fixed 25-step prior at every dose. Scheduler timesteps and initial noise scales remain family specific, so the ordered doses support within family curves, while cross family effects use codec round trip contrasts. This distinction avoids treating native strengths from different implementations as a common physical scale.

Two source specific seeds instantiate every image to image dose. SHA256 of a fixed salt, source identity, generator family, and replicate index determines each seed before detector scoring. A replicate seed remains paired across the four doses of one source and generator. Source and codec cells are deterministic and appear once. Averaging the two scheduled realizations within source defines the dose outcome while retaining source level dependence for inference. The 75,776 images are lossless 512 by 512 RGB PNG files. A versioned manifest specifies the source hash, generator revision, treatment path, dose, seed, prompt hash, and output location for every intended cell. Each generated image also carries its executed schedule and output hash. A separate scoring view contains only an intervention index, image path, and image hash. This view allows treatment blind measurement while the manifest preserves the exact crossing needed for the causal contrasts.

B.3 Detector measurements

The measurement question concerns how existing detection functions respond to the controlled paths. The panel therefore treats every detector as a frozen scalar measurement function. It contains 58 publicly available AI generated image detectors. The counting unit is one paper level method with one canonical public inference route. The count collapses alternative backbones, checkpoints, training sets, crops, and benchmark wrappers of the same method. Each method first passes an actual two image execution with its released preprocessing, weights, and score definition. The complete detector-level summary tables give every formal method name and its primary reference once.

Every detector supplies a frozen scalar outcome whose orientation assigns larger values to stronger generated image evidence. A predeclared monotone map places unbounded native logits on $[0, 1]$, while already bounded released scores retain their native scale. Raw scores remain available as a separate table. The effect magnitude is interpreted within each detector's frozen output scale; cross detector summaries use direction and interval classifications because the score units have detector specific meanings. The experiment estimates changes in these fixed measurements and introduces

no classification cutoff. MDMF and RealStats use a 1,024 image source only reference bank that is disjoint from the analysis roster, reserve roster, pilot, and the 14,534-image exclusion set.

B.4 Analysis family

The primary analysis connects each scientific question to a predeclared contrast. Detector scoring is treatment blind. The analysis joins scores to intervention metadata by the opaque identifier, averages the two seed realizations inside each source and dose, and then forms every contrast within source. The primary family contains 232 codec round trip effects, 928 source relative dose effects, 928 incremental path contrasts, 696 adjacent dose effects, and 348 cross-generator codec contrasts. These counts total 3,132 and follow directly from the 58 by 4 detector and generator crossing.

Every primary point effect is an exact average for the frozen roster, and its cluster standard error is descriptive. Under A3 and the additional A4(i) i.i.d. source interpretation, Proposition 1 gives a radius of 0.107 for every two cell contrast in the complete family. We report these intervals as assumption conditional sensitivity analyses. The signed effect vectors carry the roster conclusions. The coordinates D_j, V_j, H_j summarize codec round trip magnitude, within family dose variation, and cross-generator heterogeneity without replacing their component contrasts.

The pilot contains 256 different source units, FLUX.1 dev, and the same detector interfaces. It provides implementation checks and variance estimates. Its 1,536 images and 58 detector score vectors remain separate from all confirmatory estimates.

Two additional analyses examine precisely the information that source matching and the codec cell contribute. A prespecified incomplete design instantiates the missing codec coordinate in Corollary 1. It removes every FLUX.1 codec cell while retaining the corresponding source and image to image cells. The target family and two candidate codec means, 0.25 and 0.75, are fixed before detector outcomes are examined. Both candidates lie in the bounded score model and reconstruct the same source relative dose effects while inducing distinct codec and incremental path components. The construction changes only the observed coordinate set and uses the same detector scores as the complete experiment.

B.5 Observational content imbalance

The second analysis asks how an unmatched content mixture changes the apparent path effect. Before confirmatory scores are available, an exact token rule partitions the frozen source captions into 865 sources that mention a person term and 1,183 sources without a person term. The labels describe caption wording only. For FLUX.1 dev at dose 0.8, we compare the identified source-matched effect $\widehat{m}_j(0.8)$ with two unmatched score differences:

$$\widehat{A}_j = \overline{Y}_j(\text{img2img}, 0.8 \mid C = 1) - \overline{Y}_j(\text{source} \mid C = 0), \tag{14}$$

$$\widehat{B}_j = \overline{Y}_j(\text{img2img}, 0.8 \mid C = 0) - \overline{Y}_j(\text{source} \mid C = 1). \tag{15}$$

The same intervention and score function enter all three quantities, while Eqs. 14 and 15 use different source populations across their two terms. Their variation illustrates the observational ambiguity in Theorem 2(a). This descriptive demonstration stays outside the 3,132-member primary family.

B.6 Content conditioned effects

Source matching also makes content variation an estimable effect modifier. The same preintervention caption rule defines two conditional versions of the source-matched FLUX.1 dose 0.8 effect. Within each stratum, every retained source contributes both its source score and the mean of its two scheduled img2img scores. The analysis therefore averages a within source contrast over 865 person mention captions and 1,183 captions without a person term. Crossing the two strata with 58 detectors gives a separate family of 116 conditional effects. Proposition 1 uses the corresponding stratum size for each effect and applies one union bound across this family. Differences between the two stratum estimates are reported as descriptive summaries.

Table 5: Identification and A3/A4(i)-conditional simultaneous interval summary for the predeclared causal estimands. Roster point estimates are exact finite-roster summaries.

Estimand	Required observed cells	Target	ID	Missing	Positive	Negative	Includes zero
Codec round trip effect	Source and codec cells for the same source and generator	232	232	0	119	8	105
Source-relative dose effect	Source and image-to-image cells at the target dose	928	928	0	578	28	322
Incremental path contrast	Codec and image-to-image cells at the target dose	928	928	0	279	19	630
Adjacent dose increment	Image-to-image cells at two adjacent doses	696	696	0	43	10	643
Cross-generator codec round trip contrast	Codec cells for two generators on the same source	348	348	0	58	64	226

C Results

C.1 Confirmatory causal effects

The first question is whether the complete crossing supplies every coordinate required by the theory. The completed design contains $2{,}048$ sources, 75,776 intervention images, and 58 fixed detector measurements per image. The crossed design identifies all 3132 predeclared primary causal contrasts across 4 generator families. Every point estimate is an exact frozen roster mean. Conditional on A3 and A4(i), the simultaneous intervals exclude zero for 127 of 232 codec effects, 606 of 928 source relative dose effects, 298 of 928 incremental path effects, 53 of 696 adjacent dose effects, and 122 of 348 cross-generator codec contrasts. Table 5 reports the complete identification and sign partitions.

This complete identification matters because each contrast answers a different question. A source relative effect records the change between an input and the end of a path. A codec effect locates the change already present after the registered round trip. An incremental contrast measures what the remaining source-conditioned path adds to that codec output. Adjacent dose and cross-generator contrasts then describe how these components vary with intervention strength and implementation. The table therefore represents a decomposition of endpoint responses with distinct scientific meanings.

C.2 Codec round trips and incremental paths

The second question is which stage of the path moves detector scores. Fig. 3 displays the complete detector by generator decomposition. Among the 232 codec round trip effects, the exact roster estimates contain 207 positive, 25 negative, and 0 zero values. The corresponding A3/A4(i)-conditional interval counts are 119 above zero, 8 below zero, and 105 including zero.

The 928 incremental img2img-over-codec effects contain 695 positive, 233 negative, and 0 zero roster estimates. Their conditional interval counts are 279 above zero, 19 below zero, and 630 including zero. The largest codec effect is 0.987 for PROBE under Kolors, and the largest incremental effect is 0.753 for FatFormer under Kolors at dose 0.6. Complete signed values appear in Section G.

The identity $m = \tau^{\mathrm{dec}} + \eta$ makes the scientific meaning of this separation visible. Matching its components within each of the 928 triples defined by a detector, generator, and dose gives 261 exact opposite-sign $(\tau^{\mathrm{dec}}, \eta)$ pairs and 0 pairs with at least one exact-zero component. The A3/A4(i)-conditional simultaneous intervals resolve both components away from zero in opposite directions for 24 triples. Under the fixed algebraic definition $|m| < \min\{|\tau^{\mathrm{dec}}|, |\eta|\}$ within an exact opposite-sign pair, 50 triples exhibit strict cancellation and 0 fall on the boundary. The maximum computed identity residual $|m - \tau^{\mathrm{dec}} - \eta|$ is 3.43×10^{-15}.

FreqDetect under Kandinsky 2.2 at dose 0.8 shows what an endpoint comparison conceals. Its codec round trip moves the score by -0.197, while the incremental source-conditioned path moves it by 0.286. The complete source relative effect is only 0.088. Both components are resolved in opposite directions, so the smaller endpoint response reflects cancellation between two substantial effects. Reading the endpoint alone as a weak detector response would miss both the negative codec

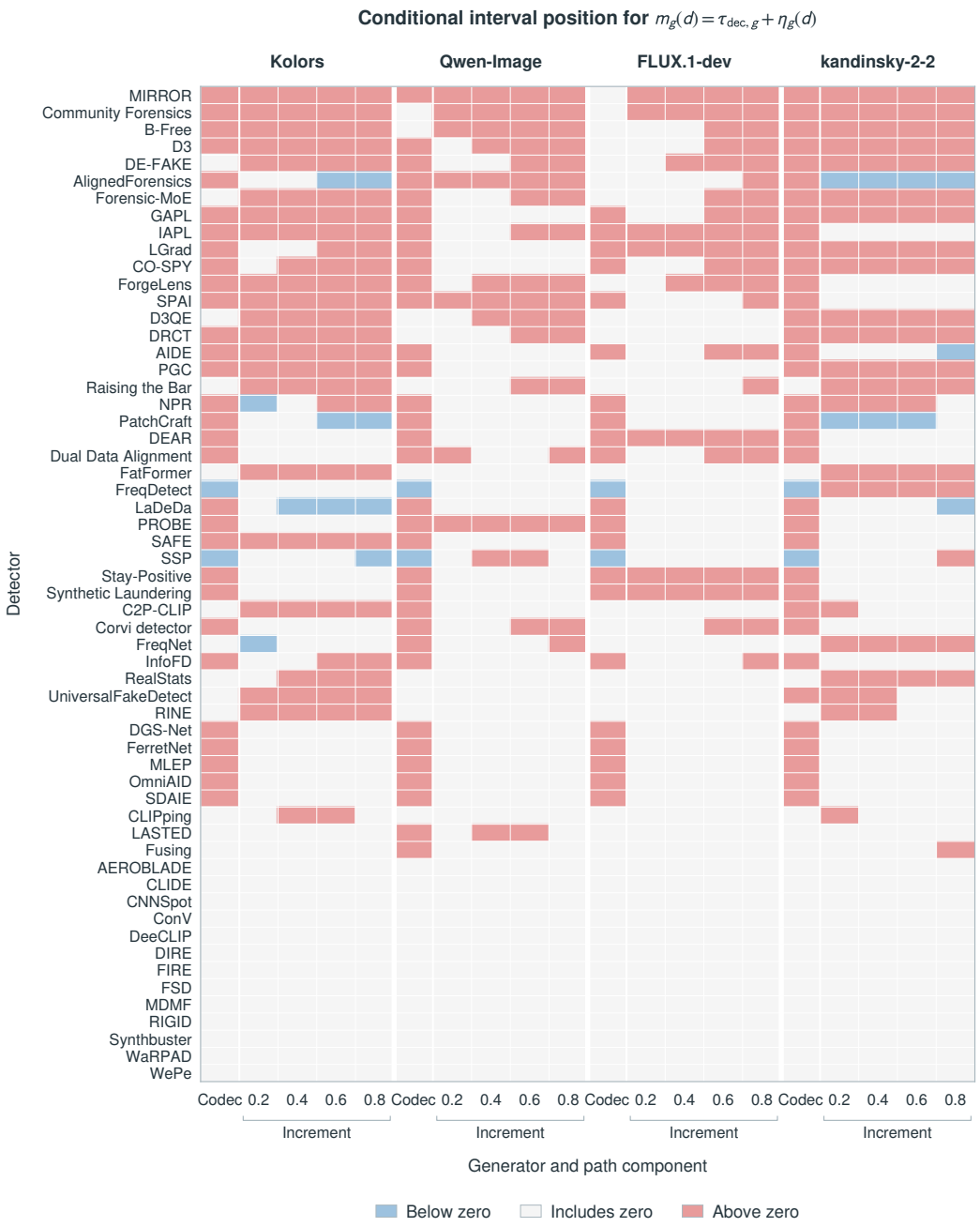

Figure 3: Complete codec round trip and incremental path responses across 58 detectors and four generator families. Colors encode A3/A4(i)-conditional interval position relative to zero; the underlying exact roster estimates are available in the accompanying tables.

sensitivity and the positive incremental sensitivity. This example directly instantiates the theoretical decomposition: the intermediate codec cell changes the scientific interpretation even though the source and final image cells remain the same.

C.3 Detector response structure

The third question is whether detectors respond to a common part of the path. Fig. 2 summarizes the fraction of codec and incremental contrasts whose conditional intervals exclude zero. At least one codec effect is supported for 40 detectors, at least one incremental effect for 40, and cross-generator

Table 6: Detector-level simultaneous evidence by adjacent dose pattern. Mixed support requires resolved changes in both directions across the 12 adjacent dose contrasts; global pattern unresolved covers every remaining complete pattern. Codec, Incr., and Hetero. count detectors with at least one resolved codec, incremental, or cross-generator contrast. The D, V, and H columns are descriptive finite-roster coordinates.

Adjacent dose pattern	n	**Codec**	**Incr.**	**Hetero.**	D	V	H
Mixed support	4	2	4	2	0.054	0.084	0.039
Global pattern unresolved	54	38	36	33	0.366	0.032	0.110

codec heterogeneity for 35. These counts show that sensitivity to both stages is widespread across the panel, while the cross-generator count shows that the same nominal stage can affect a fixed score differently across implementations.

Each detector supplies 12 adjacent dose intervals from four generators and three adjacent dose pairs. Fifty three of 696 adjacent dose intervals exclude zero, with 43 above zero and 10 below zero. The increasing and decreasing detector classes are empty. Four detectors contain supported adjacent changes in both directions, and the remaining 54 are interval inconclusive across their 12 within family comparisons. Table 6 reports these counts with the descriptive D, V, H coordinates. Thus larger native image to image strength produces no common monotone movement of detector scores. Dose is informative within a registered family, and its effect pattern remains detector specific.

Three profiles clarify the distinction among path components. FerretNet has four codec effects from 0.966 to 0.983, all with intervals above zero, while all 16 incremental intervals include zero. The experiment therefore resolves FerretNet's response at the codec stage and leaves the incremental stage unresolved on these paths. FatFormer has all four codec intervals including zero and eight of 16 incremental intervals above zero, placing its resolved response after the codec round trip. AlignedForensics exhibits a third pattern: its codec effects are 0.012 under FLUX.1 dev and 0.982 under Kandinsky 2.2, and the corresponding FLUX.1 dev minus Kandinsky contrast is -0.971. Its codec sensitivity depends strongly on the registered generator implementation. Together, these profiles show that a path-level account requires component-specific effects beyond one aggregate score shift.

C.4 What incomplete comparisons leave unresolved

The final question is whether ordinary incomplete comparisons can recover the same path interpretation. Globally hiding one codec coordinate preserves each source-relative dose effect while opening the ambiguity direction in Corollary 1. Two constructive witnesses fix the hidden codec mean at 0.2500 and 0.7500. The first gives codec effects from -0.7500 to 0.2499 and incremental effects from -0.2499 to 0.7500; the second gives codec effects from -0.2500 to 0.7499 and incremental effects from -0.7499 to 0.2500. Both reconstruct every observable source relative effect with maximum error 1.11×10^{-16}. The witnesses therefore agree on the endpoint while assigning different values to its codec and incremental components. Their endpoint agreement supplies no basis for choosing one decomposition. This is the empirical form of the row space criterion in Theorem 2: once the codec coordinate is hidden, the component contrast lies outside the observed coordinate span. The complete design closes this ambiguity by measuring that coordinate for every source and generator.

The source-caption partition supplies a separate observational illustration. The two unmatched mixtures reverse the sign of the matched FLUX.1 dose 0.8 effect for 4 and 3 of the 58 detectors. Their median deviations from the matched effect are -0.001 and 0.001. Fig. 4 and Table 7 report every detector.

The median observational deviations are small across the complete detector panel, yet selected detectors cross zero when the content allocation changes. Thus a small panel average can accompany a changed scientific conclusion for an individual method. Source matching removes this between group content term from every primary contrast and turns the comparison into a difference between interventions on the same source unit.

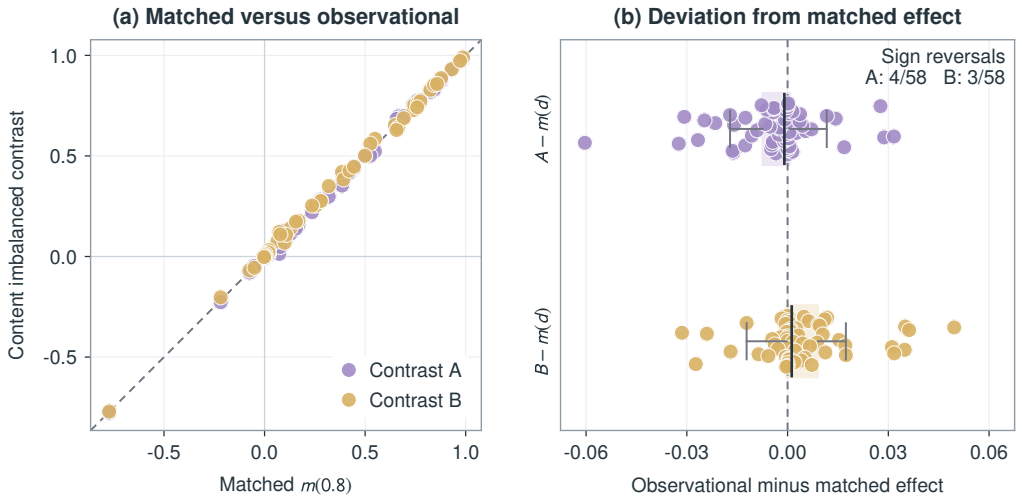

Figure 4: Content allocation changes detector responses. Left: source-matched FLUX.1 dose 0.8 effects compared with two observational contrasts whose source groups have opposite caption-content mixtures. Right: detector-level signed deviations from the matched effect; points show detectors, boxes show medians and interquartile ranges, and annotations count sign reversals.

Table 7: Practical content-imbalance demonstration at $d = 0.8$. Matched effects average the two scheduled seeds within source. A contrasts person-stratum generated scores with nonperson source scores; B reverses the content allocation.

Quantity	Mean	Median	RMS
Matched $m(d)$	0.295	0.162	0.473
Observational A	0.292	0.147	0.472
Observational B	0.300	0.176	0.475
$A - m(d)$	-0.004	-0.001	0.015
$B - m(d)$	0.004	0.001	0.015

Table 8: Conditional source-matched effects by predeclared caption stratum. Point summaries are exact for the frozen strata; CI counts use the separate A3/A4(i)-conditional family.

Stratum	n	Median	IQR	Sign $+/0/-$	CI $+/0/-$
Caption mentions person	865	0.1562	0.6560	47/0/11	30/26/2
No person term	1183	0.1666	0.6438	47/0/11	31/25/2

C.5 Content conditioned path effects

Equation 9 becomes directly measurable because the person and nonperson caption strata are fixed before detector scoring and each source retains the complete intervention crossing. Table 8 reports the two stratum summaries for the FLUX.1 dose 0.8 source-matched effect. The detector-level values and their descriptive stratum differences appear in Table 17.

The median conditional effects are 0.1562 for captions that mention a person and 0.1666 for captions without a person term. Across detectors, the median person minus nonperson difference is 0.0003, with an interquartile range of 0.0115. The aggregate similarity of the two strata supports stability for this particular predeclared content split, while the conditional estimands retain detector-level effect modification as an observable quantity. This analysis complements the imbalance demonstration: matching identifies each stratum effect, whereas mixing different source populations changes the estimand itself.

D IMPLICATIONS

A detector score becomes scientifically more useful when it is treated as a measurement of a controlled path. The detector remains fixed, and the image construction changes one registered component at a time. Under this view, the score is the outcome of an intervention instead of an explanation by itself. The codec round trip effect measures how much evidence appears when an image passes through the generator's image processor, latent representation, and decoder. The incremental contrast measures the additional movement associated with the remaining source-conditioned path. Their sum recovers the ordinary source relative response. This decomposition gives a precise answer to a question that endpoint accuracy leaves open: which part of a versioned generation route changes the detector's evidence?

The complete cell design consequently changes how a detection benchmark is interpreted. A conventional real versus generated comparison evaluates a full route and combines source composition, codec passage, conditioning, and generation into one contrast. High performance can coexist with several causal explanations for the underlying score difference. Source crossing removes the between image content comparison, and the intermediate codec cell separates two observed portions of the route. Ordered doses and paired seeds then expose within family response shape while preserving dependence across interventions. The resulting benchmark describes a detector's response to controlled operations as well as its ability to separate endpoint image collections.

The missing codec construction shows why intermediate cells carry scientific information. Once that coordinate is hidden, multiple codec and incremental effects reproduce the same source relative response exactly. More endpoint samples improve precision for the observed total, yet the component ambiguity remains. Benchmark design therefore benefits from recording images at meaningful intermediate stages, preserving source identities across every stage, and versioning the implementations that produce those stages. For the effects studied here, the complete 21 coordinate mean vector is a minimal design. This result provides a concrete planning rule: collect the cell needed for the desired contrast before expanding the number of endpoint examples.

Cancellation has a similarly direct consequence. A modest source relative effect can arise from a modest response throughout the path, or from large codec and incremental responses with opposite signs. The FreqDetect example exhibits the second pattern, and the broader panel contains many opposing component pairs. Endpoint summaries can therefore understate how strongly a detector reacts to individual operations. Reporting the signed component vector alongside the total makes this distinction visible. Simultaneous intervals add a useful layer by separating algebraic opposition in the frozen roster from cases where both directions remain resolved under the stated source sampling model.

Heterogeneity across detectors and generators further changes the meaning of a single benchmark ranking. FerretNet, FatFormer, and AlignedForensics respond to different portions of the tested paths. FerretNet shows a resolved codec response with incremental effects left unresolved, FatFormer shows the reverse pattern, and AlignedForensics changes sharply across generator codecs. These profiles describe distinct measurement behaviors even when two methods have a similar endpoint score difference. The cross-generator contrasts also show that a generic label such as codec passage is too coarse for empirical prediction. The loaded processor, latent rule, decoder, and postprocessor form a versioned intervention whose effect can vary across implementations.

These observations suggest a response oriented approach to detector design. When codec effects dominate, a detector may place substantial weight on traces introduced before source-conditioned generation. Such a response can be useful for detecting images that traverse that codec, while its transfer depends on the codec implementation. When incremental effects dominate, the measurement is more sensitive to the later path represented by the registered treatment. Opposing effects indicate that one component can suppress evidence introduced by another. Training and evaluation can use these component profiles to select augmentations, construct path specific objectives, and identify cases where combining responses with different sensitivities produces a more stable measurement. The causal design supplies the target for this work without requiring a new classification threshold.

The same framework also clarifies the level of the resulting claims. Effect magnitudes live on each detector's frozen score scale, so direction and interval status support comparisons across methods more naturally than raw magnitude. The interventions identify effects of the registered path cells on

those scores. Attribution to semantic change, frequency artifacts, latent geometry, or another internal mediator requires additional interventions that measure or manipulate the proposed mediator. Such mediator designs extend the present path-level effects to semantic, frequency, and latent mechanisms.

Taken together, complete source crossing turns a detector panel into an instrument for studying generative paths. It reveals where score evidence enters, where later operations amplify or cancel it, and where the response depends on a particular generator implementation. A benchmark built around these questions evaluates both endpoint detection and the causal structure of the evidence that produces it.

E Proofs and Assumption Details

E.1 Formal conditions

A1: well-defined intervention cells. Each $t \in \mathcal{T}$ fixes the generator identity, checkpoint, prompt rule, image dimensions, latent initialization rule, scheduler, dose, and replicate seed schedule. The condition excludes hidden versions of a treatment. Its empirical check is an immutable manifest row for every realized cell.

A2: consistency. When source z receives replicate r of cell t, the observed detector score equals the corresponding scheduled potential score:

$$Y_{ijr}^{\mathrm{obs}}(t) = Y_{jr}(z_i, t). \tag{16}$$

The image hash and intervention identifier bind the observed score to this cell.

A3: no interference. The potential score for one source-cell pair is invariant to cells instantiated for other sources. Generation services therefore use fixed state and a batching rule independent of previously generated images.

A4: target roster or source sampling. Two interpretations are available. Under A4(i), the superpopulation case, $Z_1, \ldots, Z_n$ are independent draws from a declared distribution P_Z. Under A4(ii), the finite-roster case, P_Z is the uniform empirical distribution on the frozen set $\{z_1, \ldots, z_n\}$:

$$P_Z^{\mathrm{roster}} = \frac{1}{n} \sum_{i=1}^{n} \delta_{z_i}. \tag{17}$$

The second interpretation supports conclusions for that roster. Extension to a wider source population requires a sampling argument.

A5: cell coverage. Every source contains every cell needed by its target contrast. The complete design observes the full vector

$$\mathbf{Y}_{ij} = \left(\widetilde{Y}_{ij}(t) : t \in \mathcal{T} \right). \tag{18}$$

A global omission defines an incomplete design and invokes Theorem 2. A sporadic generation or scoring failure violates coverage and remains explicit.

A6: frozen replicate schedule. The complete seed schedule is a deterministic function of preintervention source identity, generator family, and replicate index. It is fixed before any detector outcome is available. A replicate seed is paired across doses within one source and generator. The resulting scheduled cell is part of the intervention definition.

A7: frozen measurement. The detector implementation, checkpoint, scalar score, orientation, and any bounded mapping are fixed before confirmatory outcomes are examined. Scoring receives an opaque image identifier and image bytes. Treatment metadata enters only at the analysis join.

A8: moments and bounds. Identification requires

$$\mathbb{E}\big[|\overline{Y}_j(Z, t)|\big] < \infty \qquad \text{for every } j, t. \tag{19}$$

Proposition 1 further requires known finite bounds $A_j \le Y_{jr}(z, t) \le B_j$.

Interpretation conditions. A1 through A8 concern total cell effects. More specific language uses five additional checks. First, a declared content equivalence relation supports source-content fidelity. Second, decoder and image to image execution traces support path modularity. Third, recorded initialization times and update counts support within-family dose ordering. Fourth, cross-family dose comparisons use an external alignment map. Fifth, fidelity-based inclusion uses measurements fixed independently of detector j's score. These conditions affect interpretation while leaving the total cell contrast well defined.

E.2 Full cell minimality statement

Fix a detector and an unrestricted cell mean model. The joint vector containing every τ_{jg}^{dec} and $\eta_{jg}(d_{gk})$ is identified from coordinate cell means exactly when the source, every codec, and every image to image cell are observed. Thus the $1 + |\mathcal{G}| + \sum_g K_g$ scheduled cell means form a minimal coordinate design. If L_j maps the complete p dimensional cell mean vector to these components, then

$$\ker(L_j) = \mathrm{span}\{\mathbf{1}_p\}, \qquad \mathrm{rank}(L_j) = p - 1. \tag{20}$$

Suppose the source and all image to image means are observed while codec means are missing for q families. Under the variation independent mean model, the compatible component vectors contain a relatively open q dimensional patch whose affine hull is a q flat. For each missing family, a shift a_g acts as

$$\tau_{jg}^{\mathrm{dec}}(a_g) = \tau_{jg}^{\mathrm{dec}} + a_g, \qquad \eta_{jg}(d; a_g) = \eta_{jg}(d) - a_g, \tag{21}$$

and leaves every $m_{jg}(d)$ unchanged.

E.3 Seed replication

For an image to image cell, let $Y_{ijr}(t) = Y_{jr}(z_i, t)$, where the R_t seeds follow the frozen replicate schedule. The scheduled unit response is the exact arithmetic mean

$$\widetilde{Y}_{ij}(t) = \frac{1}{R_t} \sum_{r=1}^{R_t} Y_{jr}(z_i, t) = \overline{Y}_j(z_i, t). \tag{22}$$

Equation 22 defines the primary estimand for the two realized seeds. A separate random-seed experiment can instead target an expectation over a declared seed distribution. If every score lies in $[A_j, B_j]$, its replicate mean satisfies

$$A_j \le \widetilde{Y}_{ij}(t) \le B_j. \tag{23}$$

Thus replication leaves the range used by Proposition 1 unchanged.

The source-level variance of the scheduled-replicate individual contrast is a different object from the coordinate V_j:

$$V_{j;t,t'}^{\mathrm{src}} = \mathrm{Var}_Z\left[\overline{Y}_j(Z, t) - \overline{Y}_j(Z, t')\right]. \tag{24}$$

Two seeds reveal variation within the scheduled pair, while inference about a larger seed distribution requires an additional random-seed design.

E.4 Proof of Theorem 1

Proof. Full crossing supplies an observed outcome for each coefficient in Eq. 5. By consistency and Eq. 22, the source-level contrast equals

$$\sum_{t \in \mathcal{T}} c_t \widetilde{Y}_{ij}(t) = \sum_{t \in \mathcal{T}} c_t \overline{Y}_j(Z_i, t). \tag{25}$$

Taking expectation over Z_i and applying linearity yields

$$\mathbb{E}\left[\widehat{\theta}_{j,c}\right] = \sum_{t \in \mathcal{T}} c_t \mathbb{E}_Z\left[\overline{Y}_j(Z, t)\right] = \sum_{t \in \mathcal{T}} c_t \mu_j(t) = \theta_{j,c}. \tag{26}$$

Hence the estimand is a functional of the observed crossed-vector distribution, which establishes identification and unbiasedness.

Under superpopulation sampling, A4(i) and A8 make the source-level contrast vectors independent and identically distributed with finite first moment under the common frozen intervention rule. The strong law of large numbers gives

$$\widehat{\theta}_{j,c} \xrightarrow{\text{a.s.}} \theta_{j,c}. \tag{27}$$

For a fixed roster, averaging Eq. 25 over the uniform distribution in Eq. 17 gives the roster mean exactly. This proves the final statement. □

For a discrete content stratum $C = c_0$, let $N_{c_0} = \sum_i \mathbf{1}\{C_i = c_0\}$. When $N_{c_0} > 0$, the estimator

$$\widehat{\theta}_{j,\ell|c_0} = \frac{1}{N_{c_0}} \sum_{i:C_i=c_0} \sum_{t\in\mathcal{T}} \ell_t \widetilde{Y}_{ij}(t) \tag{28}$$

is conditionally unbiased. Positive stratum probability makes N_{c_0} diverge under superpopulation sampling, so the same strong-law argument establishes consistency.

For a finite predeclared family of conditional contrasts, condition on the realized preintervention stratum indicators and require every used stratum to be nonempty. Under A3, A4(i), and the bounds of Proposition 1, the observations within each stratum are independent draws from its conditional source law. The proposition therefore applies with M equal to the total declared family size and with n replaced, for each contrast, by its stratum count N_{c_0}.

Corollary 2 (Randomized incomplete assignment)**.** *Suppose cell t is observed with indicator $R_i(t)$, assignment probability $\pi_t(Z_i) = \Pr\{R_i(t) = 1 \mid Z_i\} > 0$, and $R_i(t)$ is conditionally independent of all potential outcomes given Z_i. Then*

$$\mu_j(t) = \mathbb{E}\left[\frac{R_i(t)\overline{Y}_j(Z_i,t)}{\pi_t(Z_i)}\right]. \tag{29}$$

The numerator is observed by consistency when $R_i(t) = 1$ and equals zero when $R_i(t) = 0$.

Proof. Consistency and conditional exchangeability give

$$\mathbb{E}\left[\frac{R_i(t)\overline{Y}_j(Z_i,t)}{\pi_t(Z_i)} \,\middle|\, Z_i\right] = \overline{Y}_j(Z_i,t). \tag{30}$$

Taking expectation over Z_i proves Eq. 29; linearity identifies any contrast whose cells have positive assignment probability. □

E.5 Proof of Theorem 2

Proof. For part (a), let $C, T, Y \in \{0,1\}$, with $C \sim \text{Bernoulli}(1/2)$, $T = C$, and observed $Y = T$. Model A sets $Y_A(t) = t$, while Model B sets $Y_B(t) = C$. Their causal effects are

$$\mathbb{E}[Y_A(1) - Y_A(0)] = 1, \qquad \mathbb{E}[Y_B(1) - Y_B(0)] = 0. \tag{31}$$

Both models satisfy consistency and induce the same observed law

$$\Pr(C = T = Y = 0) = \Pr(C = T = Y = 1) = \frac{1}{2}. \tag{32}$$

The observed association therefore leaves the causal effect unidentified.

For part (b), first consider a cell $t^\star$ absent for a source set B with $P_Z(B) > 0$. Choose two feasible outcome values $L < U$. Construct Model L and Model U to agree on all observed potential outcomes, and set the missing potential outcome on B to L and U, respectively. For any contrast with $c_{t^\star} \neq 0$, the two target values differ by

$$c_{t^\star} P_Z(B)(U - L) \neq 0. \tag{33}$$

Their observed laws coincide, which proves nonidentification.

The row-space criterion gives the exact mean-level statement. If $c \in \mathrm{row}(P_\Omega)$, there is a vector a such that $c = P_\Omega^\top a$, and

$$c^\top \mu = a^\top P_\Omega \mu \tag{34}$$

is observed. Conversely, if $c \notin \mathrm{row}(P_\Omega)$, finite-dimensional orthogonality provides $v \in \ker(P_\Omega)$ with $c^\top v \neq 0$. For a sufficiently small ϵ, choose an interior point $\mu^\circ \in \mathcal{M}$. Both μ° and $\mu^\circ + \epsilon v$ are feasible and satisfy

$$P_\Omega(\mu^\circ + \epsilon v) = P_\Omega \mu^\circ, \qquad c^\top(\mu^\circ + \epsilon v) \neq c^\top \mu^\circ. \tag{35}$$

This proves necessity. If the decoder mean is missing, any feasible $r = \mu_j(t_g^{\mathrm{dec}})$ yields

$$\tau_{jg}^{\mathrm{dec}}(r) = r - \mu_j(t_0), \qquad \eta_{jg}(d; r) = \mu_j(t_{gk}^{\mathrm{i2i}}) - r, \qquad \tau_{jg}^{\mathrm{dec}}(r) + \eta_{jg}(d; r) = m_{jg}(d). \tag{36}$$

The retained cells determine the last line for every r, while the first two lines vary.

For part (c), begin with any observed potential score process $Y(t)$ and use the same response map $h(m, s) = m + s$ in both models. Mediator Model M sets $M(t) = Y(t)$ and $S(t) = 0$. Mediator Model S sets $M(t) = 0$ and $S(t) = Y(t)$. Both imply

$$h(M(t), S(t)) = Y(t) \tag{37}$$

for every cell. One model assigns the full score response to M, and the other assigns it to S. Their treatment and score distributions coincide. This construction operates in the unrestricted latent-mediator class; component semantics require measured mediators and additional structural or exclusion restrictions. □

Proof of Corollary 1. Complete observation is sufficient by Theorem 1. For necessity, the row space of a coordinate selector contains only coefficient vectors that vanish on every missing coordinate. If the source mean is missing, each codec round trip effect has a nonzero source coefficient. If decoder coordinate g is missing, both its codec effect and each incremental contrast in that family have a nonzero coefficient on that mean. If an image to image mean is missing, its corresponding incremental contrast has a nonzero coefficient on that mean. The row-space criterion in Theorem 2(b) therefore makes every listed coordinate necessary.

For Eq. 20, $L_j v = 0$ first implies $v(t_g^{\mathrm{dec}}) = v(t_0)$ from every decoder component and then $v(t_{gk}^{\mathrm{i2i}}) = v(t_g^{\mathrm{dec}})$ from every incremental component. Hence all entries of v are equal. Constant vectors plainly lie in the kernel, so rank-nullity gives rank $p - 1$.

Now let G_{mis} contain the q families with missing decoder means, and let e_g be the standard basis vector of decoder coordinate g. For $g \in G_{\mathrm{mis}}$, varying only $\mu_j(t_g^{\mathrm{dec}})$ gives a null direction of the observation map. Its image in the component vector adds one to the coordinate τ_{jg}^{dec}, subtracts one from every $\eta_{jg}(d_{gk})$, and leaves all other family components fixed. These q images are linearly independent because each contains a distinct decoder effect coordinate. Variation independence supplies a compatible μ° whose missing decoder coordinates lie in their interval interiors and positive ϵ_g for which the component vectors contain the patch

$$\left\{ L_j \mu^\circ + \sum_{g \in G_{\mathrm{mis}}} a_g L_j e_g : a_g \in (-\epsilon_g, \epsilon_g) \right\}. \tag{38}$$

Its affine hull is a q-flat. Equation 12 proves invariance of every total dose response. □

Corollary 3 (Missing dose)**.** *An absent dose mean and its adjacent increments are unidentified from the remaining ordered doses unless a shape restriction supplies additional information.*

Proof. Apply Theorem 2(b) with $t^\star$ equal to the missing dose cell. Dose order alone places no numeric restriction on its mean. □

E.6 Proof of Proposition 1

Let $\mathcal{H}$ contain M predeclared contrasts. Their source level values, targets, and estimators are

$$X_{ih} = \sum_{j,t} c_{h,j,t} \widetilde{Y}_{ij}(t), \qquad \theta_h = \mathbb{E}_{P_Z}[X_{ih}], \qquad \widehat{\theta}_h = \frac{1}{n} \sum_{i=1}^{n} X_{ih}. \tag{39}$$

Proposition 1 (Bounded simultaneous intervals under source sampling)**.** *Under A4(i), suppose every score lies in its frozen interval $[A_j, B_j]$ and source clusters are independent. For $n \geq 1$ and $0 < \alpha < 1$,*

$$R_h = \sum_{j,t} |c_{h,j,t}|(B_j - A_j), \qquad r_h(\alpha) = R_h \sqrt{\frac{\log(2M/\alpha)}{2n}}, \qquad \Pr\Big\{ |\widehat{\theta}_h - \theta_h| \leq r_h(\alpha) \; \forall h \in \mathcal{H} \Big\} \geq 1 - \alpha. \tag{40}$$

Under A4(ii), the same estimator equals the finite-roster target exactly.

Proof. For contrast h, each score component lies in its detector-specific bounded interval. The source-level contrast therefore has range width at most

$$\sup X_{ih} - \inf X_{ih} \leq \sum_{j,t} |c_{h,j,t}|(B_j - A_j) = R_h. \tag{41}$$

If $R_h = 0$, Eq. 41 makes X_{ih} constant almost surely, and the coverage event holds with radius zero. Consider $R_h > 0$ for the remaining calculation.

Under A4(i), independence holds across source clusters; dependence inside a source cluster remains unrestricted. Hoeffding's inequality and Eq. 40 give

$$\Pr\Big(|\widehat{\theta}_h - \theta_h| > r_h(\alpha) \Big) \leq 2 \exp\left(-\frac{2n r_h(\alpha)^2}{R_h^2} \right) = \frac{\alpha}{M}. \tag{42}$$

A union bound over $\mathcal{H}$ yields

$$\Pr\left(\bigcup_{h \in \mathcal{H}} \{ |\widehat{\theta}_h - \theta_h| > r_h(\alpha) \} \right) \leq \sum_{h \in \mathcal{H}} \frac{\alpha}{M} = \alpha. \tag{43}$$

Taking complements proves Eq. 40. The argument allows arbitrary dependence among all measurements inside source i. □

E.7 Empirical identification checks

Assumptions A1, A5 through A7, and the bounded score component of A8 have direct empirical checks. Every experimental unit specifies the generator revision, treatment path, dose, replicate index, seed, source hash, and output hash. Each detector observation supplies the detector identity, image hash, scalar score, and intervention index. These quantities establish a one to one correspondence between each scheduled treatment and its measured outcome. Image to image cells contain exactly two replicate indices, and the same replicate seed follows a source and generator pair across doses. A2 consistency, A3 noninterference, and A4(i) source sampling remain interpretive assumptions. A4(ii) defines the finite-roster target directly.

Before aggregation, the design requires the complete detector by intervention crossing. Equation 4 gives the within source replicate mean, Eq. 39 gives each source level contrast, and the reported coefficient vector identifies its cell combination. Replicate uniqueness, score completeness, stable source identity, and scheduled cell consistency are necessary for the corresponding estimate. A decoder cell absent for every source remains a valid incomplete design and yields the nonidentification result in Eq. 12.

F Detector Methods and Public Resources

Table 9 lists every evaluated method, its primary reference, and its public code or model repository.

Table 9: Publicly available detector methods, primary references, and code or model repositories.

Method	Reference	Code or model repository
AEROBLADE	(Ricker et al., 2024)	jonasricker/aeroblade
AIDE	(Yan et al., 2025b)	sadjadeb/DetectZoo
AlignedForensics	(Rajan et al., 2025)	AniSundar18/AlignedForensics
B-Free	(Guillaro et al., 2025)	grip-unina/B-Free
C2P-CLIP	(Tan et al., 2025)	sadjadeb/DetectZoo
CLIDE	(Betser et al., 2026)	FujitsuResearch/domain-adaptive-image-detection
CLIPping	(Khan & Dang-Nguyen, 2024)	sfimediafutures/CLIPping-the-Deception
CNNSpot	(Wang et al., 2020)	PeterWang512/CNNDetection
CO-SPY	(Cheng et al., 2025)	sadjadeb/DetectZoo
Community Forensics	(Park & Owens, 2025)	JeongsooP/Community-Forensics
ConV	(Zhang et al., 2025b)	tmlr-group/ConV
Corvi detector	(Corvi et al., 2023)	grip-unina/DMimageDetection
D3	(Yang et al., 2025)	sadjadeb/DetectZoo
D3QE	(Zhang et al., 2025a)	Zhangyr2022/D3QE
DE-FAKE	(Sha et al., 2023)	mever-team/sidbench
DEAR	(Kim et al., 2026)	dahyedahye/dear
DeeCLIP	(Keita et al., 2025)	Mamadou-Keita/DeeCLIP
DGS-Net	(Yan et al., 2025a)	HorizonTEL/DGS-Net
DIRE	(Wang et al., 2023)	ZhendongWang6/DIRE
DRCT	(Chen et al., 2024a)	sadjadeb/DetectZoo
Dual Data Alignment	(Chen et al., 2025a)	roy-ch/Dual-Data-Alignment
FatFormer	(Liu et al., 2024)	sadjadeb/DetectZoo
FerretNet	(Liang et al., 2025)	xigua7105/FerretNet
FIRE	(Chu et al., 2025)	Chuchad/FIRE
Forensic-MoE	(Fang et al., 2025)	fangmq77/Forensic-MoE
ForgeLens	(Chen et al., 2025b)	Yingjian-Chen/ForgeLens
FreqDetect	(Frank et al., 2020)	mever-team/sidbench
FreqNet	(Tan et al., 2024b)	sadjadeb/DetectZoo
FSD	(Nguyen et al., 2025)	ductai199x/Forensic-Self-Descriptions-CVPR25
Fusing	(Ju et al., 2022)	mever-team/sidbench
GAPL	(Qin et al., 2026)	UltraCapture/GAPL
IAPL	(Li et al., 2026)	liyih/IAPL
InfoFD	(Qin et al., 2025)	Ant0ny44/InfoFD
LaDeDa	(Cavia et al., 2024)	sadjadeb/DetectZoo
LASTED	(Wu et al., 2025b)	HighwayWu/LASTED
LGrad	(Tan et al., 2023)	sadjadeb/DetectZoo
MDMF	(Zhang et al., 2026)	ZBox1005/MDMF
MIRROR	(Liu et al., 2026)	349793927/MIRROR
MLEP	(Yuan et al., 2025)	fkeufss/MLEP
NPR	(Tan et al., 2024a)	chuangchuangtan/NPR-DeepfakeDetection
OmniAID	(Guo et al., 2026)	yunncheng/OmniAID
PatchCraft	(Zhong et al., 2023)	sadjadeb/DetectZoo
PGC	(Zhou et al., 2026)	xiaoyu6868/PGC
PROBE	(Cao et al., 2026)	Amamiya-C/PROBE-AIGI-Detection
Raising the Bar	(Cozzolino et al., 2024)	grip-unina/ClipBased-SyntheticImageDetection
RealStats	(Zisman & Shaham, 2026)	shaham-lab/RealStats
RIGID	(He et al., 2024)	IBM/RIGID
RINE	(Koutlis & Papadopoulos, 2024)	mever-team/sidbench
SAFE	(Li et al., 2025)	sadjadeb/DetectZoo
SDAIE	(Zhong et al., 2026)	Ekko-zn/SDAIE
SPAI	(Karageorgiou et al., 2025)	mever-team/spai
SSP	(Chen et al., 2024b)	bcmi/SSP-AI-Generated-Image-Detection
Stay-Positive	(Rajan & Lee, 2025)	AniSundar18/AlignedForensics
Synthbuster	(Bammey, 2024)	qbammey/synthbuster
Synthetic Laundering	(Mandelli et al., 2024)	polimi-ispl/synthetic-image-detection
UniversalFakeDetect	(Ojha et al., 2023)	WisconsinAIVision/UniversalFakeDetect
WaRPAD	(Choi et al., 2025)	sungikchoi/WaRPAD
WePe	(Nie et al., 2025)	tmlr-group/WePe

G Complete Confirmatory Contrasts

The following tables report the exact frozen-roster point effects for every detector and generator. Table 10 defines the generator keys. Tables 11, 12, 13, and 14 give the four complete path-effect matrices. Tables 15 and 16 report detector coordinates and cross-generator codec round trip contrasts.

H Complete Conditional Content Effects

Table 17 reports the complete detector panel for the two preintervention caption strata. Each conditional effect is source-matched within its stratum. The final column gives the descriptive difference between the two frozen stratum means.

Table 10: Generator identities used in the complete confirmatory result tables.

Code	Family	Generator identity
G1	kolors	`Kwai-Kolors/Kolors-diffusers`
G2	qwen-image	`Qwen/Qwen-Image`
G3	flux-1	`black-forest-labs/FLUX.1-dev`
G4	kandinsky-2-2	`kandinsky-community/kandinsky-2-2-decoder`

Table 11: Path effects for G1. Values are finite-roster estimates rounded to three decimals. Red, blue, and black denote A3/A4(i) intervals above zero, below zero, and including zero. I, D, M, and U denote increasing, decreasing, mixed, and unresolved dose response.

Detector	τ	$m_{.2}$	$m_{.4}$	$m_{.6}$	$m_{.8}$	$\eta_{.2}$	$\eta_{.4}$	$\eta_{.6}$	$\eta_{.8}$	Class
AEROBLADE	.003	.003	.004	.005	.006	1.2e−4	.001	.002	.003	U
AIDE	.529	.725	.752	.743	.711	.197	.223	.214	.182	U
AlignedForensics	.943	.971	.911	.738	.737	.029	−.032	−.205	−.205	U
B-Free	.875	.989	.990	.990	.991	.114	.115	.116	.116	U
C2P-CLIP	.017	.215	.462	.576	.494	.199	.445	.559	.477	U
CLIDE	1.1e−6	3.0e−5	6.5e−5	8.1e−5	7.5e−5	2.9e−5	6.4e−5	8.0e−5	7.4e−5	U
CLIPping	−.019	.079	.149	.136	.062	.098	.169	.155	.081	U
CNNSpot	3.9e−4	−4.2e−4	−.001	−.003	−.003	−8.1e−4	−.001	−.003	−.003	U
Community Forensics	.625	.996	.998	.999	.999	.371	.373	.374	.374	U
ConV	.002	.010	.013	.012	.008	.008	.011	.010	.005	U
Corvi detector	.855	.869	.827	.792	.863	.014	−.028	−.063	.008	U
CO-SPY	.379	.304	.571	.686	.677	−.074	.193	.307	.299	U
D3	.218	.704	.842	.876	.843	.486	.623	.657	.624	U
D3QE	.070	.452	.567	.513	.361	.382	.497	.444	.291	M
DEAR	.746	.763	.779	.781	.790	.017	.033	.035	.044	U
DeeCLIP	.003	.037	.058	.064	.056	.034	.054	.060	.053	U
DE-FAKE	.083	.245	.473	.638	.683	.162	.390	.555	.600	U
DGS-Net	.885	.934	.937	.939	.936	.049	.053	.054	.051	U
DIRE	2.8e−6	−1.7e−4	3.1e−5	−2.4e−5	−.002	−1.7e−4	2.8e−5	−2.7e−5	−.002	U
DRCT	.114	.338	.482	.607	.646	.224	.368	.493	.532	U
Dual Data Alignment	.868	.937	.938	.938	.938	.069	.070	.070	.070	U
FatFormer	.026	.563	.775	.779	.650	.537	.748	.753	.624	M
FerretNet	.974	.984	.984	.986	.983	.010	.010	.012	.009	U
FIRE	−1.1e−7	1.9e−6	1.3e−6	2.0e−6	−8.1e−7	2.0e−6	1.5e−6	2.1e−6	−7.0e−7	U
Forensic-MoE	.101	.332	.521	.591	.507	.231	.420	.489	.405	U
FSD	.004	−1.9e−4	−.007	−.013	−.014	−.004	−.011	−.018	−.018	U
ForgeLens	.159	.594	.763	.825	.820	.435	.604	.666	.661	U
FreqDetect	−.149	−.206	−.194	−.181	−.161	−.057	−.045	−.031	−.012	U
FreqNet	.072	−.072	−.031	.021	.057	−.144	−.103	−.051	−.015	U
Fusing	.001	−.004	−.006	−.004	−.002	−.006	−.007	−.005	−.004	U
GAPL	.787	.955	.965	.966	.966	.167	.178	.179	.179	U
IAPL	.294	.727	.845	.874	.865	.433	.551	.579	.571	U
InfoFD	.160	.229	.254	.276	.281	.069	.094	.116	.121	U
LaDeDa	.162	.067	.015	.017	.002	−.095	−.147	−.145	−.160	U
LASTED	.014	.074	.056	.058	.062	.060	.042	.044	.048	U
LGrad	.117	.044	.202	.292	.287	−.073	.085	.175	.170	U
MDMF	2.6e−5	.002	.003	.003	.002	.002	.003	.003	.002	U
MIRROR	.344	.887	.892	.893	.893	.543	.548	.548	.549	U
MLEP	.500	.528	.533	.534	.535	.028	.033	.034	.035	U
NPR	.324	.189	.389	.476	.488	−.134	.065	.152	.165	U
OmniAID	.771	.815	.813	.801	.789	.045	.043	.030	.019	U
PatchCraft	.895	.915	.838	.769	.764	.020	−.057	−.126	−.131	U
PGC	.253	.635	.676	.674	.647	.382	.424	.421	.395	U
PROBE	.987	.994	.994	.994	.994	.008	.008	.008	.008	U
Raising the Bar	.044	.258	.395	.479	.520	.213	.351	.435	.475	U
RealStats	−.052	−.041	.068	.105	.057	.011	.120	.157	.109	U
RIGID	.001	.004	.009	.010	.008	.003	.008	.009	.007	U
RINE	.009	.366	.632	.587	.304	.357	.622	.577	.295	M
SAFE	.720	.832	.859	.855	.857	.112	.139	.134	.137	U
SDAIE	.973	.973	.965	.938	.937	3.7e−4	−.008	−.035	−.036	U
SPAI	.479	.779	.780	.750	.725	.300	.301	.270	.246	U
SSP	−.637	−.710	−.704	−.707	−.762	−.073	−.067	−.071	−.125	U
Stay-Positive	.661	.637	.604	.573	.581	−.023	−.057	−.088	−.080	U
Synthbuster	1.8e−7	2.1e−7	2.1e−7	2.3e−7	2.2e−7	2.9e−8	3.8e−8	5.5e−8	4.0e−8	U
Synthetic Laundering	.764	.702	.679	.688	.701	−.062	−.085	−.077	−.064	U
UniversalFakeDetect	.043	.294	.401	.318	.161	.251	.358	.275	.119	M
WaRPAD	8.2e−4	.002	.002	.002	.002	7.8e−4	9.1e−4	.001	9.6e−4	U
WePe	9.1e−4	.004	.007	.006	.002	.004	.006	.005	.001	U

Table 12: Path effects for G2. Values are finite-roster estimates rounded to three decimals. Red, blue, and black denote A3/A4(i) intervals above zero, below zero, and including zero. I, D, M, and U denote increasing, decreasing, mixed, and unresolved dose response.

Detector	τ	$m_{.2}$	$m_{.4}$	$m_{.6}$	$m_{.8}$	$\eta_{.2}$	$\eta_{.4}$	$\eta_{.6}$	$\eta_{.8}$	Class
AEROBLADE	.002	.002	.002	.002	.003	4.3e−4	8.7e−5	1.3e−5	.001	U
AIDE	.623	.660	.656	.651	.684	.036	.033	.028	.060	U
AlignedForensics	.292	.623	.572	.510	.531	.332	.281	.218	.239	U
B-Free	.104	.521	.577	.612	.637	.417	.473	.508	.533	U
C2P-CLIP	.273	.225	.280	.324	.285	−.048	.007	.051	.012	U
CLIDE	2.2e−6	3.1e−6	1.1e−6	3.1e−6	2.1e−5	8.7e−7	−1.1e−6	8.3e−7	1.9e−5	U
CLIPping	.076	.082	.091	.091	.049	.005	.015	.015	−.027	U
CNNSpot	4.5e−4	.002	.002	.003	−.001	.002	.001	.003	−.002	U
Community Forensics	.082	.295	.300	.334	.582	.212	.218	.252	.499	U
ConV	.003	.003	.003	.003	.001	−2.8e−4	−3.6e−5	−6.6e−5	−.002	U
Corvi detector	.816	.893	.909	.930	.958	.077	.092	.114	.141	U
CO-SPY	.723	.725	.717	.716	.781	.002	−.006	−.007	.058	U
D3	.370	.421	.497	.575	.648	.050	.126	.204	.277	U
D3QE	.060	.124	.179	.202	.186	.064	.119	.143	.126	U
DEAR	.856	.847	.847	.850	.860	−.009	−.010	−.006	.003	U
DeeCLIP	.056	.061	.062	.059	.051	.005	.005	.002	−.005	U
DE-FAKE	.144	.177	.236	.362	.585	.033	.092	.218	.441	U
DGS-Net	.922	.931	.933	.933	.932	.009	.010	.010	.010	U
DIRE	1.8e−5	−.002	−.003	−.002	−.001	−.002	−.003	−.002	−.001	U
DRCT	.002	.052	.086	.153	.247	.049	.084	.151	.244	U
Dual Data Alignment	.691	.818	.789	.785	.841	.127	.098	.094	.150	U
FatFormer	.035	.068	.082	.094	.073	.033	.047	.058	.037	U
FerretNet	.966	.958	.959	.966	.977	−.008	−.006	6.0e−4	.011	U
FIRE	−1.1e−5	−6.6e−7	−2.0e−6	−3.4e−6	−3.8e−5	1.0e−5	8.6e−6	7.3e−6	−2.7e−5	U
Forensic-MoE	.443	.449	.493	.586	.627	.006	.050	.143	.184	U
FSD	.005	.005	.005	.005	.005	−3.0e−5	3.9e−4	4.5e−4	5.5e−4	U
ForgeLens	.591	.612	.711	.809	.859	.021	.121	.218	.268	U
FreqDetect	−.201	−.201	−.203	−.204	−.182	−3.5e−5	−.002	−.003	.019	U
FreqNet	.393	.421	.416	.431	.537	.028	.023	.038	.144	U
Fusing	.125	.200	.168	.152	.158	.075	.043	.027	.033	U
GAPL	.738	.819	.815	.799	.845	.080	.077	.061	.107	U
IAPL	.664	.720	.742	.786	.824	.056	.079	.122	.160	U
InfoFD	.225	.235	.257	.288	.306	.010	.031	.062	.081	U
LaDeDa	.535	.461	.450	.454	.458	−.074	−.085	−.081	−.077	U
LASTED	.119	.212	.297	.296	.176	.092	.178	.177	.057	U
LGrad	.411	.469	.474	.457	.445	.058	.063	.046	.034	U
MDMF	5.0e−6	2.1e−5	4.5e−5	8.6e−5	9.7e−5	1.6e−5	4.0e−5	8.1e−5	9.2e−5	U
MIRROR	.322	.728	.774	.801	.854	.406	.452	.479	.532	U
MLEP	.524	.528	.528	.527	.526	.004	.004	.004	.002	U
NPR	.495	.501	.504	.504	.505	.006	.009	.009	.010	U
OmniAID	.592	.654	.649	.638	.656	.062	.057	.047	.065	U
PatchCraft	.943	.943	.946	.947	.952	−4.5e−4	.003	.004	.009	U
PGC	.289	.316	.336	.369	.373	.028	.047	.081	.084	U
PROBE	.435	.680	.637	.647	.777	.245	.202	.212	.342	U
Raising the Bar	.037	.066	.106	.163	.279	.028	.069	.126	.242	U
RealStats	−.043	−.052	−.045	−.031	.002	−.009	−.003	.012	.045	U
RIGID	3.4e−4	8.6e−4	3.4e−4	1.3e−4	3.4e−4	5.2e−4	−2.5e−6	−2.1e−4	1.6e−6	U
RINE	.011	.006	.009	.018	.023	−.005	−.002	.006	.011	U
SAFE	.662	.709	.716	.726	.715	.047	.054	.064	.053	U
SDAIE	.973	.973	.971	.970	.966	−2.4e−4	−.002	−.003	−.007	U
SPAI	.286	.520	.603	.672	.677	.234	.316	.385	.391	U
SSP	−.569	−.481	−.457	−.454	−.520	.087	.111	.115	.048	U
Stay-Positive	.683	.683	.683	.689	.699	−8.3e−4	−3.7e−4	.005	.016	U
Synthbuster	−3.6e−5	−9.6e−5	−1.1e−4	−1.1e−4	−1.1e−4	−6.0e−5	−7.6e−5	−7.0e−5	−7.0e−5	U
Synthetic Laundering	.600	.639	.654	.681	.698	.039	.054	.081	.098	U
UniversalFakeDetect	.061	.066	.093	.105	.061	.006	.032	.045	8.0e−5	U
WaRPAD	.002	.001	.001	.001	.001	−4.2e−4	−4.9e−4	−3.7e−4	−3.3e−4	U
WePe	1.4e−4	−1.3e−4	−4.3e−4	−6.7e−4	−.002	−2.7e−4	−5.7e−4	−8.0e−4	−.002	U

I Text to Image Route Comparison

I.1 Complete route estimands

The complete route comparison adds one text to image cell per generator. Define

$$\mathcal{T}^{+} = \mathcal{T} \cup \{t_g^{\mathrm{T2I}} : g \in \mathcal{G}\}, \tag{44}$$

where t_g^{T2I} uses the frozen source caption and the two scheduled integer seeds. Let d_{gK_g} be the highest registered image to image dose. The text to image source relative effect and complete route contrast are

Table 13: Path effects for G3. Values are finite-roster estimates rounded to three decimals. Red, blue, and black denote A3/A4(i) intervals above zero, below zero, and including zero. I, D, M, and U denote increasing, decreasing, mixed, and unresolved dose response.

Detector	τ	$m_{.2}$	$m_{.4}$	$m_{.6}$	$m_{.8}$	$\eta_{.2}$	$\eta_{.4}$	$\eta_{.6}$	$\eta_{.8}$	Class
AEROBLADE	8.0e−4	.001	.002	.002	.003	4.4e−4	8.3e−4	.001	.002	U
AIDE	.450	.476	.515	.570	.648	.026	.065	.120	.198	U
AlignedForensics	.012	.020	.036	.072	.129	.009	.025	.060	.118	U
B-Free	.055	.143	.149	.195	.255	.088	.093	.139	.199	U
C2P-CLIP	.023	.014	.019	.042	.064	−.009	−.004	.020	.042	U
CLIDE	−2.5e−7	2.4e−6	7.0e−6	1.6e−5	2.7e−5	2.7e−6	7.3e−6	1.6e−5	2.7e−5	U
CLIPping	−.036	−.041	−.041	−.047	−.075	−.006	−.005	−.012	−.039	U
CNNSpot	5.8e−4	−6.0e−4	−4.7e−4	−8.0e−4	−.002	−.001	−.001	−.001	−.002	U
Community Forensics	.080	.230	.324	.453	.739	.150	.243	.373	.659	U
ConV	6.7e−4	−1.6e−4	−2.9e−4	−3.2e−4	−.002	−8.3e−4	−9.6e−4	−9.9e−4	−.002	U
Corvi detector	.012	.080	.105	.130	.169	.068	.093	.118	.156	U
CO-SPY	.486	.527	.566	.622	.744	.041	.080	.136	.258	U
D3	.096	.071	.114	.228	.385	−.025	.019	.132	.289	U
D3QE	.016	.056	.083	.102	.073	.040	.067	.086	.057	U
DEAR	.457	.583	.630	.697	.764	.125	.173	.239	.306	U
DeeCLIP	.005	.002	.002	.004	.005	−.004	−.003	−.001	−1.2e−4	U
DE-FAKE	.058	.107	.189	.307	.549	.049	.131	.249	.491	U
DGS-Net	.832	.898	.903	.916	.930	.066	.070	.084	.098	U
DIRE	−2.2e−5	1.8e−5	−3.3e−5	−3.9e−4	−2.4e−4	4.1e−5	−1.1e−5	−3.7e−4	−2.2e−4	U
DRCT	.023	−.035	−.019	.017	.100	−.058	−.043	−.006	.076	U
Dual Data Alignment	.385	.413	.430	.512	.665	.028	.045	.127	.281	U
FatFormer	.014	.008	.004	.012	.022	−.006	−.010	−.002	.008	U
FerretNet	.983	.981	.981	.983	.986	−.002	−.001	−9.2e−5	.003	U
FIRE	−5.3e−6	−1.7e−4	−3.0e−4	−1.0e−6	6.0e−7	−1.6e−4	−2.9e−4	4.3e−6	5.9e−6	U
Forensic-MoE	.039	.076	.118	.196	.280	.037	.079	.157	.240	U
FSD	.005	.005	.004	.004	.002	1.1e−4	−1.4e−4	−5.1e−4	−.002	U
ForgeLens	.078	.176	.269	.461	.692	.098	.191	.383	.614	U
FreqDetect	−.221	−.220	−.221	−.220	−.218	6.9e−4	3.0e−4	4.8e−4	.002	U
FreqNet	.030	.040	.052	.063	.107	.011	.022	.033	.078	U
Fusing	.001	.005	.007	.011	.019	.004	.006	.010	.018	U
GAPL	.543	.551	.634	.744	.878	.008	.091	.201	.335	U
IAPL	.163	.455	.537	.695	.825	.292	.373	.532	.661	U
InfoFD	.145	.186	.206	.243	.280	.041	.062	.098	.136	U
LaDeDa	.263	.269	.247	.232	.236	.006	−.017	−.032	−.027	U
LASTED	.012	.017	.014	.011	.006	.006	.002	−.001	−.006	U
LGrad	.158	.277	.295	.302	.319	.119	.137	.144	.161	U
MDMF	4.3e−6	−1.5e−5	−2.0e−5	−2.5e−5	−3.0e−5	−1.9e−5	−2.4e−5	−2.9e−5	−3.4e−5	U
MIRROR	.030	.237	.404	.563	.775	.207	.374	.534	.745	I
MLEP	.493	.508	.510	.511	.526	.015	.017	.018	.033	U
NPR	.453	.461	.467	.475	.499	.008	.014	.022	.046	U
OmniAID	.689	.779	.773	.762	.759	.090	.084	.073	.070	U
PatchCraft	.797	.853	.858	.860	.843	.056	.061	.063	.045	U
PGC	.096	.062	.058	.064	.078	−.034	−.038	−.031	−.018	U
PROBE	.358	.297	.267	.308	.392	−.061	−.092	−.050	.034	U
Raising the Bar	.010	.026	.049	.087	.156	.016	.039	.077	.146	U
RealStats	−.029	−.037	−.049	−.055	−.050	−.008	−.020	−.026	−.021	U
RIGID	2.9e−4	4.5e−4	6.7e−4	9.3e−4	.001	1.6e−4	3.8e−4	6.5e−4	.001	U
RINE	.006	−.002	−.003	−.003	−.003	−.008	−.008	−.009	−.009	U
SAFE	.814	.844	.847	.846	.858	.031	.033	.033	.044	U
SDAIE	.938	.968	.968	.970	.973	.030	.031	.033	.035	U
SPAI	.283	.370	.349	.369	.423	.087	.066	.086	.139	U
SSP	−.711	−.644	−.657	−.698	−.773	.068	.054	.013	−.061	U
Stay-Positive	.217	.343	.373	.414	.444	.126	.156	.196	.227	U
Synthbuster	−1.8e−6	−1.3e−6	−1.1e−6	−1.2e−6	5.4e−8	4.9e−7	7.1e−7	6.3e−7	1.9e−6	U
Synthetic Laundering	.428	.573	.593	.620	.657	.146	.166	.192	.229	U
UniversalFakeDetect	.014	.009	.007	.005	−9.2e−4	−.005	−.007	−.009	−.015	U
WaRPAD	2.7e−4	6.9e−5	1.2e−4	3.5e−4	4.9e−4	−2.0e−4	−1.4e−4	8.2e−5	2.3e−4	U
WePe	1.8e−4	−6.8e−4	−.001	−.001	−.002	−8.6e−4	−.001	−.001	−.003	U

$$\rho_{jg}^{\mathrm{T2I}} = \mu_j(t_g^{\mathrm{T2I}}) - \mu_j(t_0), \tag{45}$$

$$\zeta_{jg}^{\mathrm{T2I}} = \mu_j(t_g^{\mathrm{T2I}}) - \mu_j(t_{gK_g}^{\mathrm{i2i}}) = \rho_{jg}^{\mathrm{T2I}} - m_{jg}(d_{gK_g}). \tag{46}$$

Equal integer seeds preserve the scheduled pairing across the two cells. Each pipeline defines its own latent tensor and random stream.

Corollary 4 (Identification of complete route contrasts)**.** *Under A1 through A8 on $\mathcal{T}^+$, complete source crossing identifies Eqs. 45 and 46. Their source mean estimators follow Eq. 10. The supplement places the $58 \times 4 = 232$ route contrasts in a separate simultaneous family.*

Table 14: Path effects for G4. Values are finite-roster estimates rounded to three decimals. Red, blue, and black denote A3/A4(i) intervals above zero, below zero, and including zero. I, D, M, and U denote increasing, decreasing, mixed, and unresolved dose response.

Detector	τ	$m_{.2}$	$m_{.4}$	$m_{.6}$	$m_{.8}$	$\eta_{.2}$	$\eta_{.4}$	$\eta_{.6}$	$\eta_{.8}$	Class
AEROBLADE	.008	.008	.009	.009	.009	9.5e−6	4.0e−4	6.1e−4	7.2e−4	U
AIDE	.663	.706	.692	.626	.551	.044	.029	−.036	−.112	U
AlignedForensics	.982	.856	.801	.809	.818	−.127	−.181	−.173	−.165	U
B-Free	.869	.990	.989	.989	.988	.121	.120	.120	.119	U
C2P-CLIP	.213	.350	.293	.282	.286	.136	.080	.069	.073	U
CLIDE	2.6e−6	4.7e−5	5.0e−5	4.9e−5	4.9e−5	4.4e−5	4.7e−5	4.7e−5	4.7e−5	U
CLIPping	.079	.251	.164	.144	.149	.172	.084	.065	.069	U
CNNSpot	−.001	.003	−.002	2.0e−4	.002	.004	−6.2e−4	.001	.003	U
Community Forensics	.479	.996	.995	.992	.989	.517	.515	.513	.510	U
ConV	.006	.012	.007	.006	.006	.006	.001	1.7e−5	−3.0e−5	U
Corvi detector	.939	.963	.947	.933	.929	.024	.008	−.005	−.010	U
CO-SPY	.350	.836	.810	.806	.804	.486	.460	.456	.454	U
D3	.648	.911	.904	.901	.898	.263	.256	.253	.250	U
D3QE	.127	.409	.302	.273	.283	.282	.175	.146	.156	U
DEAR	.772	.825	.825	.823	.817	.052	.052	.051	.044	U
DeeCLIP	.027	.068	.060	.055	.054	.041	.032	.027	.026	U
DE-FAKE	.119	.591	.672	.682	.683	.472	.553	.563	.564	U
DGS-Net	.938	.938	.938	.934	.915	−3.4e−4	−4.6e−5	−.004	−.023	U
DIRE	−2.0e−4	−2.7e−5	−1.1e−4	−.001	−.004	1.7e−4	8.6e−5	−.001	−.003	U
DRCT	.190	.582	.618	.603	.580	.392	.428	.413	.390	U
Dual Data Alignment	.843	.934	.933	.932	.930	.091	.090	.089	.087	U
FatFormer	.034	.548	.350	.285	.291	.514	.316	.251	.257	U
FerretNet	.977	.982	.970	.964	.953	.005	−.007	−.014	−.024	U
FIRE	−2.6e−6	1.5e−6	2.6e−6	2.7e−6	2.7e−6	4.0e−6	5.1e−6	5.2e−6	5.3e−6	U
Forensic-MoE	.816	.980	.973	.970	.970	.164	.157	.154	.155	U
FSD	.004	7.7e−4	−.003	−.003	−.004	−.003	−.007	−.006	−.007	U
ForgeLens	.905	.912	.912	.911	.910	.007	.007	.006	.005	U
FreqDetect	−.197	.141	.125	.088	.088	.339	.322	.285	.286	U
FreqNet	.044	.608	.484	.443	.405	.564	.439	.398	.360	U
Fusing	.027	.112	.108	.128	.159	.085	.081	.101	.132	U
GAPL	.776	.963	.964	.966	.965	.187	.188	.190	.189	U
IAPL	.890	.901	.901	.901	.901	.011	.011	.011	.011	U
InfoFD	.328	.358	.360	.363	.362	.030	.031	.035	.034	U
LaDeDa	.393	.394	.403	.337	.272	.001	.011	−.056	−.121	U
LASTED	.053	.047	.033	.038	.035	−.006	−.020	−.015	−.018	U
LGrad	.184	.473	.427	.413	.402	.289	.243	.229	.218	U
MDMF	2.3e−5	5.7e−4	2.1e−4	1.3e−4	1.3e−4	5.5e−4	1.8e−4	1.1e−4	1.1e−4	U
MIRROR	.629	.893	.893	.893	.892	.264	.264	.264	.264	U
MLEP	.509	.533	.530	.520	.507	.024	.021	.011	−.002	U
NPR	.373	.512	.499	.486	.478	.139	.127	.113	.105	U
OmniAID	.737	.637	.662	.673	.671	−.100	−.075	−.064	−.066	U
PatchCraft	.928	.708	.759	.808	.837	−.219	−.169	−.120	−.090	U
PGC	.366	.649	.624	.586	.563	.284	.258	.220	.198	U
PROBE	.962	.994	.993	.992	.990	.032	.031	.030	.028	U
Raising the Bar	.093	.318	.404	.429	.443	.225	.311	.336	.351	U
RealStats	−.026	.156	.122	.103	.109	.182	.147	.129	.135	U
RIGID	4.4e−4	.011	.008	.007	.008	.011	.008	.007	.007	U
RINE	.006	.303	.142	.101	.092	.296	.135	.095	.086	U
SAFE	.768	.868	.860	.860	.857	.100	.092	.092	.089	U
SDAIE	.978	.910	.944	.956	.962	−.067	−.034	−.021	−.015	U
SPAI	.382	.332	.348	.336	.287	−.050	−.034	−.045	−.095	U
SSP	−.881	−.870	−.868	−.784	−.653	.011	.013	.097	.228	U
Stay-Positive	.691	.654	.646	.638	.625	−.037	−.045	−.052	−.066	U
Synthbuster	−1.0e−7	2.2e−7	2.3e−7	2.3e−7	2.3e−7	3.2e−7	3.3e−7	3.3e−7	3.3e−7	U
Synthetic Laundering	.747	.685	.671	.673	.672	−.062	−.076	−.074	−.075	U
UniversalFakeDetect	.139	.391	.248	.224	.221	.253	.109	.085	.082	U
WaRPAD	.001	.001	.002	.002	.002	2.9e−4	5.1e−4	6.4e−4	6.8e−4	U
WePe	7.3e−4	.005	.002	9.9e−4	9.0e−4	.005	.002	2.6e−4	1.7e−4	U

I.2 Supplementary design

The secondary crossing supplies two text-only realizations for each source caption and generator family. It uses the same 2048 captions, 4 versioned generator snapshots, and two scheduled numerical seeds as the primary design; the text-only pipelines receive no source pixels. Each family retains its registered guidance configuration and requested 20 step schedule, with the Kandinsky prior fixed at 25 steps. Equal numerical seeds encode the analysis pairing, while latent tensors and random streams remain specific to each pipeline implementation.

The resulting $2{,}048 \times 4 \times 2 = 16{,}384$ RGB PNG files yield 950,272 canonical and 950,272 raw detector scores. We compare each complete text-only route with the paired dose-0.8 source-conditioned route through $\zeta_{jg}^{\mathrm{T2I}}$ in Eq. 46. These 232 detector by generator contrasts form a separate

Table 15: Detector response summary. The D, V, and H columns are exact finite-roster coordinates; support entries show resolved interval counts over totals.

Detector	D	V	H	Codec	Incremental	Γ
AEROBLADE	0.0045	0.0007	0.0048	0/4	0/16	0/6
AIDE	0.5723	0.0445	0.1354	4/4	7/16	3/6
AlignedForensics	0.6963	0.0636	0.6820	3/4	11/16	5/6
B-Free	0.6193	0.0300	0.6475	2/4	14/16	4/6
C2P-CLIP	0.1738	0.0872	0.1858	2/4	5/16	4/6
CLIDE	1.80×10^{-6}	1.31×10^{-5}	1.81×10^{-6}	0/4	0/16	0/6
CLIPping	0.0586	0.0420	0.0865	0/4	3/16	2/6
CNNSpot	0.0007	0.0022	0.0011	0/4	0/16	0/6
Community Forensics	0.3979	0.1191	0.3934	2/4	16/16	5/6
ConV	0.0036	0.0022	0.0030	0/4	0/16	0/6
Corvi detector	0.7548	0.0325	0.6110	3/4	4/16	4/6
CO-SPY	0.5061	0.0953	0.2396	4/4	9/16	5/6
D3	0.3916	0.0807	0.3365	3/4	13/16	6/6
D3QE	0.0787	0.0693	0.0645	1/4	11/16	1/6
DEAR	0.7239	0.0311	0.2455	4/4	4/16	4/6
DeeCLIP	0.0314	0.0075	0.0349	0/4	0/16	0/6
DE-FAKE	0.1061	0.1402	0.0536	2/4	13/16	0/6
DGS-Net	0.8954	0.0082	0.0666	4/4	0/16	0/6
DIRE	0.0001	0.0010	0.0001	0/4	0/16	0/6
DRCT	0.1114	0.0724	0.1225	2/4	10/16	3/6
Dual Data Alignment	0.7228	0.0536	0.3144	4/4	4/16	5/6
FatFormer	0.0286	0.0938	0.0141	0/4	8/16	0/6
FerretNet	0.9750	0.0064	0.0100	4/4	0/16	0/6
FIRE	6.10×10^{-6}	9.32×10^{-5}	6.41×10^{-6}	0/4	0/16	0/6
Forensic-MoE	0.4673	0.0789	0.5059	2/4	12/16	5/6
FSD	0.0044	0.0030	0.0007	0/4	0/16	0/6
ForgeLens	0.5474	0.1130	0.5469	3/4	10/16	5/6
FreqDetect	0.1938	0.0154	0.0429	4/4	4/16	0/6
FreqNet	0.2015	0.0563	0.2445	1/4	6/16	3/6
Fusing	0.0640	0.0151	0.0834	1/4	1/16	2/6
GAPL	0.7180	0.0573	0.1614	4/4	10/16	3/6
IAPL	0.5800	0.0749	0.4723	4/4	10/16	6/6
InfoFD	0.2263	0.0224	0.1180	4/4	3/16	2/6
LaDeDa	0.3660	0.0324	0.2286	4/4	4/16	5/6
LASTED	0.0660	0.0431	0.0713	1/4	2/16	1/6
LGrad	0.2458	0.0553	0.1869	4/4	10/16	3/6
MDMF	1.77×10^{-5}	0.0005	1.63×10^{-5}	0/4	0/16	0/6
MIRROR	0.3932	0.0929	0.3462	3/4	16/16	5/6
MLEP	0.5067	0.0066	0.0187	4/4	0/16	0/6
NPR	0.4165	0.0637	0.1093	4/4	6/16	3/6
OmniAID	0.7003	0.0117	0.1103	4/4	0/16	2/6
PatchCraft	0.8926	0.0375	0.0926	4/4	5/16	2/6
PGC	0.2693	0.0238	0.1605	3/4	8/16	4/6
PROBE	0.7445	0.0487	0.4740	4/4	4/16	4/6
Raising the Bar	0.0549	0.0711	0.0486	0/4	11/16	0/6
RealStats	0.0389	0.0394	0.0173	0/4	7/16	0/6
RIGID	0.0006	0.0019	0.0005	0/4	0/16	0/6
RINE	0.0084	0.1224	0.0038	0/4	6/16	0/6
SAFE	0.7431	0.0101	0.0920	4/4	4/16	1/6
SDAIE	0.9654	0.0134	0.0262	4/4	0/16	0/6
SPAI	0.3665	0.0405	0.1316	4/4	9/16	2/6
SSP	0.7090	0.0575	0.1899	4/4	4/16	4/6
Stay-Positive	0.5975	0.0225	0.3266	4/4	4/16	3/6
Synthbuster	1.79×10^{-5}	5.14×10^{-6}	2.49×10^{-5}	0/4	0/16	0/6
Synthetic Laundering	0.6491	0.0197	0.2214	4/4	4/16	5/6
UniversalFakeDetect	0.0790	0.0747	0.0756	1/4	6/16	1/6
WaRPAD	0.0010	0.0001	0.0008	0/4	0/16	0/6
WePe	0.0006	0.0017	0.0006	0/4	0/16	0/6

simultaneous family with common conditional radius 0.0945. Corollary 4 identifies the complete route contrast under the extended source crossing.

Table 18 summarizes the primary and supplementary analysis layers.

I.3 Results

Fig. 5 plots the primary source relative effect $m_{jg}(0.8)$ against the text-only source relative effect ρ_{jg}^{T2I}. Vertical displacement from the identity line is the complete route contrast $\zeta_{jg}^{\mathrm{T2I}}$.

Table 20 reports the family level distribution and sign summaries. Across the frozen panel, 132, 0, and 100 point contrasts are respectively positive, zero, and negative. Under the separate A3/A4(i)

Table 16: All cross-generator codec round trip contrasts. Exact finite-roster point estimates are shown. A3/A4(i)-conditional interval symbols indicate intervals above zero, below zero, and including zero.

Detector	$G1:G2$	$G1:G3$	$G1:G4$	$G2:G3$	$G2:G4$	$G3:G4$
AEROBLADE	0.0007°	0.0018°	-0.0058°	0.0010°	-0.0065°	-0.0076°
AIDE	-0.0947°	0.0786°	$-0.1342^{\downarrow}$	$0.1732^{\uparrow}$	-0.0396°	$-0.2128^{\downarrow}$
AlignedForensics	$0.6511^{\uparrow}$	$0.9313^{\uparrow}$	-0.0396°	$0.2801^{\uparrow}$	$-0.6908^{\downarrow}$	$-0.9709^{\downarrow}$
B-Free	$0.7706^{\uparrow}$	$0.8193^{\uparrow}$	0.0054°	0.0488°	$-0.7651^{\downarrow}$	$-0.8139^{\downarrow}$
C2P-CLIP	$-0.2567^{\downarrow}$	-0.0062°	$-0.1966^{\downarrow}$	$0.2505^{\uparrow}$	0.0601°	$-0.1904^{\downarrow}$
CLIDE	$-1.12\times10^{-6\circ}$	$1.36\times10^{-6\circ}$	$-1.48\times10^{-6\circ}$	$2.48\times10^{-6\circ}$	$-3.62\times10^{-7\circ}$	$-2.84\times10^{-6\circ}$
CLIPping	-0.0956°	0.0164°	-0.0986°	$0.1119^{\uparrow}$	-0.0030°	$-0.1149^{\downarrow}$
CNNSpot	$-6.38\times10^{-5\circ}$	-0.0002°	0.0015°	-0.0001°	0.0015°	0.0017°
Community Forensics	$0.5426^{\uparrow}$	$0.5447^{\uparrow}$	$0.1459^{\uparrow}$	0.0021°	$-0.3967^{\downarrow}$	$-0.3988^{\downarrow}$
ConV	-0.0006°	0.0018°	-0.0034°	0.0025°	-0.0027°	-0.0052°
Corvi detector	0.0391°	$0.8432^{\uparrow}$	-0.0834°	$0.8041^{\uparrow}$	$-0.1225^{\downarrow}$	$-0.9266^{\downarrow}$
CO-SPY	$-0.3443^{\downarrow}$	$-0.1075^{\downarrow}$	0.0287°	$0.2368^{\uparrow}$	$0.3729^{\uparrow}$	$0.1361^{\uparrow}$
D3	$-0.1522^{\downarrow}$	$0.1223^{\uparrow}$	$-0.4295^{\downarrow}$	$0.2745^{\uparrow}$	$-0.2772^{\downarrow}$	$-0.5518^{\downarrow}$
D3QE	0.0101°	0.0536°	-0.0572°	0.0436°	-0.0673°	$-0.1109^{\downarrow}$
DEAR	$-0.1102^{\downarrow}$	$0.2888^{\uparrow}$	-0.0262°	$0.3989^{\uparrow}$	0.0839°	$-0.3150^{\downarrow}$
DeeCLIP	-0.0528°	-0.0019°	-0.0240°	0.0510°	0.0289°	-0.0221°
DE-FAKE	-0.0610°	0.0245°	-0.0364°	0.0855°	0.0246°	-0.0608°
DGS-Net	-0.0378°	0.0523°	-0.0536°	0.0901°	-0.0158°	-0.1059°
DIRE	$-1.48\times10^{-5\circ}$	$2.51\times10^{-5\circ}$	0.0002°	$3.99\times10^{-5\circ}$	0.0002°	0.0002°
DRCT	$0.1117^{\uparrow}$	0.0907°	-0.0761°	-0.0210°	$-0.1878^{\downarrow}$	$-0.1668^{\downarrow}$
Dual Data Alignment	$0.1772^{\uparrow}$	$0.4835^{\uparrow}$	0.0249°	$0.3063^{\uparrow}$	$-0.1523^{\downarrow}$	$-0.4586^{\downarrow}$
FatFormer	-0.0089°	0.0126°	-0.0081°	0.0215°	0.0009°	-0.0206°
FerretNet	0.0083°	-0.0086°	-0.0032°	-0.0169°	-0.0115°	0.0055°
FIRE	$1.06\times10^{-5\circ}$	$5.23\times10^{-6\circ}$	$2.47\times10^{-6\circ}$	$-5.32\times10^{-6\circ}$	$-8.09\times10^{-6\circ}$	$-2.77\times10^{-6\circ}$
Forensic-MoE	$-0.3416^{\downarrow}$	0.0623°	$-0.7142^{\downarrow}$	$0.4039^{\uparrow}$	$-0.3727^{\downarrow}$	$-0.7766^{\downarrow}$
FSD	-0.0006°	-0.0004°	0.0005°	0.0002°	0.0011°	0.0009°
ForgeLens	$-0.4315^{\downarrow}$	0.0812°	$-0.7457^{\downarrow}$	$0.5127^{\uparrow}$	$-0.3142^{\downarrow}$	$-0.8270^{\downarrow}$
FreqDetect	0.0512°	0.0715°	0.0481°	0.0203°	-0.0031°	-0.0234°
FreqNet	$-0.3208^{\downarrow}$	0.0422°	0.0275°	$0.3630^{\uparrow}$	$0.3483^{\uparrow}$	-0.0147°
Fusing	$-0.1239^{\downarrow}$	$8.95\times10^{-5\circ}$	-0.0256°	$0.1240^{\uparrow}$	0.0982°	-0.0257°
GAPL	0.0489°	$0.2445^{\uparrow}$	0.0114°	$0.1956^{\uparrow}$	-0.0375°	$-0.2331^{\downarrow}$
IAPL	$-0.3694^{\downarrow}$	$0.1309^{\uparrow}$	$-0.5959^{\downarrow}$	$0.5003^{\uparrow}$	$-0.2264^{\downarrow}$	$-0.7268^{\downarrow}$
InfoFD	-0.0654°	0.0154°	$-0.1681^{\downarrow}$	0.0808°	-0.1027°	$-0.1835^{\downarrow}$
LaDeDa	$-0.3731^{\downarrow}$	-0.1015°	$-0.2308^{\downarrow}$	$0.2716^{\uparrow}$	$0.1423^{\uparrow}$	$-0.1293^{\downarrow}$
LASTED	-0.1059°	0.0019°	-0.0395°	$0.1078^{\uparrow}$	0.0664°	-0.0414°
LGrad	$-0.2948^{\downarrow}$	-0.0414°	-0.0674°	$0.2534^{\uparrow}$	$0.2273^{\uparrow}$	-0.0261°
MDMF	$2.07\times10^{-5\circ}$	$2.15\times10^{-5\circ}$	$2.41\times10^{-6\circ}$	$7.40\times10^{-7\circ}$	$-1.83\times10^{-5\circ}$	$-1.91\times10^{-5\circ}$
MIRROR	0.0225°	$0.3145^{\uparrow}$	$-0.2846^{\downarrow}$	$0.2920^{\uparrow}$	$-0.3071^{\downarrow}$	$-0.5991^{\downarrow}$
MLEP	-0.0239°	0.0067°	-0.0093°	0.0306°	0.0145°	-0.0160°
NPR	$-0.1713^{\downarrow}$	$-0.1291^{\downarrow}$	-0.0489°	0.0422°	$0.1225^{\uparrow}$	0.0803°
OmniAID	$0.1792^{\uparrow}$	0.0821°	0.0336°	-0.0971°	$-0.1455^{\downarrow}$	-0.0485°
PatchCraft	-0.0484°	0.0975°	-0.0327°	$0.1459^{\uparrow}$	0.0157°	$-0.1303^{\downarrow}$
PGC	-0.0362°	$0.1567^{\uparrow}$	$-0.1132^{\downarrow}$	$0.1929^{\uparrow}$	-0.0769°	$-0.2699^{\downarrow}$
PROBE	$0.5517^{\uparrow}$	$0.6283^{\uparrow}$	0.0246°	0.0766°	$-0.5271^{\downarrow}$	$-0.6037^{\downarrow}$
Raising the Bar	0.0072°	0.0342°	-0.0483°	0.0270°	-0.0555°	-0.0825°
RealStats	-0.0097°	-0.0230°	-0.0264°	-0.0133°	-0.0167°	-0.0034°
RIGID	0.0007°	0.0007°	0.0006°	$5.55\times10^{-5\circ}$	-0.0001°	-0.0002°
RINE	-0.0017°	0.0040°	0.0032°	0.0056°	0.0049°	-0.0008°
SAFE	0.0582°	-0.0935°	-0.0480°	$-0.1517^{\downarrow}$	-0.1062°	0.0455°
SDAIE	-0.0001°	0.0351°	-0.0049°	0.0352°	-0.0048°	-0.0400°
SPAI	$0.1929^{\uparrow}$	$0.1959^{\uparrow}$	0.0976°	0.0030°	-0.0953°	-0.0983°
SSP	-0.0681°	0.0746°	$0.2440^{\uparrow}$	$0.1427^{\uparrow}$	$0.3122^{\uparrow}$	$0.1695^{\uparrow}$
Stay-Positive	-0.0227°	$0.4436^{\uparrow}$	-0.0299°	$0.4663^{\uparrow}$	-0.0072°	$-0.4735^{\downarrow}$
Synthbuster	$3.59\times10^{-5\circ}$	$2.01\times10^{-6\circ}$	$2.77\times10^{-7\circ}$	$-3.39\times10^{-5\circ}$	$-3.56\times10^{-5\circ}$	$-1.73\times10^{-6\circ}$
Synthetic Laundering	$0.1647^{\uparrow}$	$0.3368^{\uparrow}$	0.0175°	$0.1721^{\uparrow}$	$-0.1472^{\downarrow}$	$-0.3193^{\downarrow}$
UniversalFakeDetect	-0.0179°	0.0290°	-0.0960°	0.0469°	-0.0780°	$-0.1249^{\downarrow}$
WaRPAD	-0.0007°	0.0006°	-0.0003°	0.0013°	0.0004°	-0.0008°
WePe	0.0008°	0.0007°	0.0002°	$-3.89\times10^{-5\circ}$	-0.0006°	-0.0006°

conditional sensitivity analysis, the interval counts are 25 above zero, 7 below zero, and 200 including zero; the common radius is 0.0945. These counts describe differences between the two versioned complete routes. Separate interventions are required to identify source conditioning, denoising, or random stream components.

The two route profiles remain strongly aligned across detectors, with descriptive Pearson correlations of 0.977, 0.978, 1.000, and 0.996 for FLUX.1, Qwen Image, Kandinsky 2.2, and Kolors. The median

Table 17: Detector-level conditional source-matched effects. The final column is a derived descriptive difference without an interval.

Detector	$m \mid C_{\text{person}}$	$m \mid C_{\text{nonperson}}$	Person minus nonperson
AEROBLADE	0.0030	0.0031	-0.0001
AIDE	0.6536	0.6447	0.0090
AlignedForensics	0.1129	0.1408	-0.0279
B-Free	0.2535	0.2557	-0.0022
C2P-CLIP	0.0445	0.0787	-0.0342
CLIDE	0.0000	0.0000	0.0000
CLIPping	-0.0701	-0.0786	0.0085
CNNSpot	-0.0007	-0.0022	0.0015
Community Forensics	0.7567	0.7267	0.0300
ConV	-0.0020	-0.0016	-0.0004
Corvi detector	0.1557	0.1779	-0.0222
CO-SPY	0.7621	0.7308	0.0313
D3	0.3947	0.3780	0.0167
D3QE	0.0332	0.1025	-0.0693
DEAR	0.7569	0.7685	-0.0117
DeeCLIP	0.0046	0.0056	-0.0011
DE-FAKE	0.5844	0.5232	0.0612
DGS-Net	0.9312	0.9295	0.0017
DIRE	-0.0000	-0.0004	0.0004
DRCT	0.0897	0.1071	-0.0173
Dual Data Alignment	0.6935	0.6443	0.0492
FatFormer	0.0210	0.0221	-0.0011
FerretNet	0.9887	0.9845	0.0042
FIRE	-0.0000	0.0000	-0.0000
Forensic-MoE	0.2817	0.2780	0.0037
FSD	0.0037	0.0013	0.0024
ForgeLens	0.6955	0.6897	0.0058
FreqDetect	-0.1950	-0.2355	0.0404
FreqNet	0.1340	0.0879	0.0460
Fusing	0.0157	0.0221	-0.0064
GAPL	0.9025	0.8604	0.0421
IAPL	0.8215	0.8269	-0.0054
InfoFD	0.2825	0.2786	0.0038
LaDeDa	0.2408	0.2328	0.0080
LASTED	0.0058	0.0056	0.0002
LGrad	0.3547	0.2932	0.0615
MDMF	-0.0000	-0.0000	0.0000
MIRROR	0.7772	0.7726	0.0045
MLEP	0.5637	0.4983	0.0653
NPR	0.5042	0.4956	0.0086
OmniAID	0.7492	0.7661	-0.0170
PatchCraft	0.8376	0.8461	-0.0084
PGC	0.0818	0.0755	0.0063
PROBE	0.4032	0.3838	0.0194
Raising the Bar	0.1568	0.1554	0.0014
RealStats	-0.0577	-0.0448	-0.0129
RIGID	0.0012	0.0014	-0.0002
RINE	0.0001	-0.0053	0.0054
SAFE	0.8641	0.8528	0.0113
SDAIE	0.9726	0.9729	-0.0004
SPAI	0.4146	0.4284	-0.0138
SSP	-0.7811	-0.7663	-0.0148
Stay-Positive	0.4372	0.4497	-0.0125
Synthbuster	0.0000	-0.0000	0.0000
Synthetic Laundering	0.6577	0.6556	0.0021
UniversalFakeDetect	-0.0029	0.0005	-0.0035
WaRPAD	0.0006	0.0004	0.0001
WePe	-0.0032	-0.0020	-0.0011

Table 18: Primary and supplementary analysis layers.

Analysis layer	New images	Score rows	Targets
Primary path effects, $M = 3{,}132$	75,776	4,395,008	3,132
Conditional effects, $M = 116$	0	0	116
Text to image routes, $M = 232$	16,384	950,272	232

absolute complete route contrast across all 232 pairs is 0.007, and 32 simultaneous intervals exclude zero. FLUX.1 and Qwen Image contribute 17 and 13 of these resolved contrasts, Kandinsky 2.2 contributes zero, and Kolors contributes two. The largest positive contrast is 0.378 for Community Forensics under Qwen Image; the largest negative contrast is -0.348 for SSP under the same

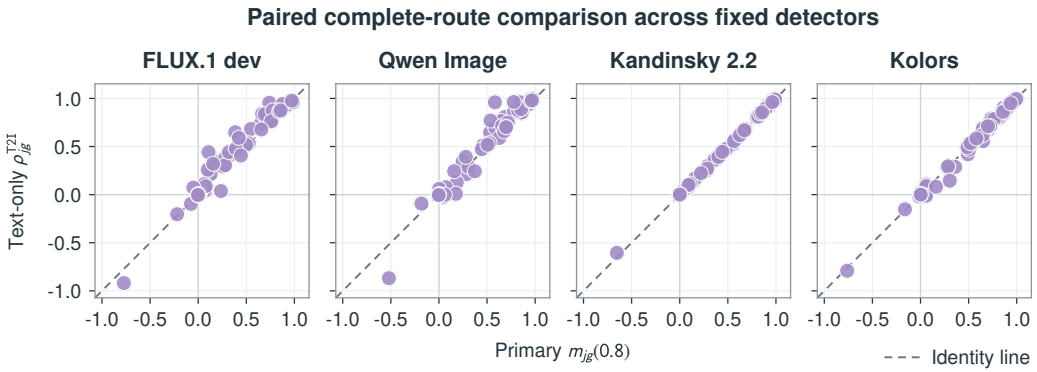

Figure 5: Paired complete route comparison for all 58 fixed detectors and 4 generator families. Each panel compares the primary dose 0.8 source relative effect with its text-only counterpart. The dashed line denotes equality of the two complete routes.

Table 19: Text to image route extension by generator family. Correlation compares detector profiles; resolved signs count route contrast intervals.

Generator	Profile correlation	Median route effect	Resolved $+/-$
FLUX.1 dev	0.977	0.0088	15/2
Qwen Image	0.978	0.0073	10/3
Kandinsky 2.2	1.000	0.0000	0/0
Kolors	0.996	-0.0003	0/2

Table 20: Complete text-only versus dose-0.8 source-conditioned route contrasts. Point summaries are exact for the frozen roster; interval counts are A3/A4(i)-conditional sensitivity summaries.

Family	Median	IQR	Min/max	Sign $+/0/-$	CI $+/0/-$
FLUX.1 dev	0.0088	0.0964	[-0.1986, 0.3342]	43/0/15	15/41/2
Qwen Image	0.0073	0.0594	[-0.3484, 0.3776]	38/0/20	10/45/3
Kandinsky 2.2	-0.0000	0.0047	[-0.0254, 0.0478]	27/0/31	0/58/0
Kolors	-0.0003	0.0152	[-0.1587, 0.0530]	24/0/34	0/56/2

generator. The broad profile agreement and these concentrated deviations show that route choice preserves much of the detector response structure while creating substantial method specific changes in selected cases. MIRROR has a resolved FLUX.1 route increase of 0.100, while its other three route intervals include zero.

Table 19 separates profile correlation, median route effect, and resolved signs by generator family.

J Complete Text to Image Supplementary Contrasts

Tables 21, 22, 23, and 24 report the complete detector by generator results. Each row gives the exact frozen-roster point effect, the descriptive source-cluster standard error, and the separate A3/A4(i)-conditional simultaneous interval.

Each complete table fixes one generator family and follows the same 58-detector order. For detector j and family g, the first numerical column gives the primary source-conditioned effect $m_{jg}(0.8)$, and the second gives the text-only effect ρ^{T2I}_{jg}. Their difference is the complete route contrast $\zeta^{\mathrm{T2I}}_{jg} = \rho^{\mathrm{T2I}}_{jg} - m_{jg}(0.8)$. Consequently, a positive value means that the text-only route produces a larger detector response than the paired source-conditioned route. A negative value means that the text-only response is smaller. This row-level subtraction keeps the detector, source caption, generator revision, and scheduled numerical seed aligned across the two complete routes.

The point estimate and the simultaneous interval answer complementary questions. The point estimate is the exact mean contrast for the frozen source roster. The source-cluster standard error summarizes sampling variability across the matched source units. The final three columns give the A3/A4(i)-conditional simultaneous interval and its sign classification within the separate family of 232 route contrasts. An interval above zero resolves a larger text-only response, an interval below zero resolves a smaller response, and an interval including zero leaves the direction unresolved under the stated source-sampling interpretation. The common interval construction makes these classifications comparable across all detector and generator rows.

Family-level profile correlation and detector-level route contrasts summarize different aspects of the same comparison. Correlation measures whether the 58 detectors retain a similar ordering and broad response geometry across the two routes. The contrast $\zeta_{jg}^{\mathrm{T2I}}$ measures the displacement of one detector within that profile. A correlation near one can therefore coexist with several resolved contrasts when most detectors preserve their relative positions while selected detectors move. The family summaries report this global agreement, and Tables 21 through 24 expose the corresponding detector-level changes.

The four generator families instantiate distinct registered pipelines, so each table retains its own implementation-specific interpretation. Native guidance, latent construction, scheduler behavior, and decoder configuration belong to the intervention definition for that family. Cross-family comparisons therefore focus on recurring detector response patterns, interval counts, and the spread of route contrasts. Within a family, every row follows one common intervention rule and supports direct detector-level comparison.

A compact reading order connects the tables to the theory. The median and interquartile range first describe the center and spread of each family. The resolved sign counts identify how many detectors exhibit route-sensitive responses under simultaneous uncertainty. The extreme rows then locate the methods that contribute most strongly to the family spread. Together, these summaries connect the extended treatment set in Corollary 4 to the complete numerical evidence below.

Table 21: Complete FLUX.1 dev text to image route contrasts. Point estimates are exact frozen-roster summaries, SE is the descriptive source-cluster standard error, and intervals are A3/A4(i)-conditional.

Detector	Family	$m(0.8)$	ρ^{T2I}	ζ^{T2I}	SE	Lower	Upper	CI status
AEROBLADE	FLUX.1 dev	0.0031	0.0059	0.0028	0.0000	-0.0917	0.0972	includes 0
AIDE	FLUX.1 dev	0.6484	0.7338	0.0854	0.0052	-0.0091	0.1798	includes 0
AlignedForensics	FLUX.1 dev	0.1290	0.2148	0.0858	0.0047	-0.0086	0.1803	includes 0
B-Free	FLUX.1 dev	0.2547	0.2927	0.0379	0.0076	-0.0565	0.1324	includes 0
C2P-CLIP	FLUX.1 dev	0.0643	0.1098	0.0456	0.0039	-0.0489	0.1400	includes 0
CLIDE	FLUX.1 dev	0.0000	0.0000	0.0000	0.0000	-0.0945	0.0945	includes 0
CLIPping	FLUX.1 dev	-0.0750	-0.0957	-0.0207	0.0019	-0.1152	0.0738	includes 0
CNNSpot	FLUX.1 dev	-0.0016	-0.0025	-0.0009	0.0008	-0.0954	0.0935	includes 0
Community Forensics	FLUX.1 dev	0.7394	0.9556	0.2162	0.0068	0.1218	0.3107	> 0
ConV	FLUX.1 dev	-0.0018	-0.0041	-0.0023	0.0001	-0.0968	0.0921	includes 0
Corvi detector	FLUX.1 dev	0.1685	0.3416	0.1730	0.0061	0.0786	0.2675	> 0
CO-SPY	FLUX.1 dev	0.7440	0.8506	0.1066	0.0053	0.0122	0.2011	> 0
D3	FLUX.1 dev	0.3851	0.6481	0.2630	0.0089	0.1686	0.3575	> 0
D3QE	FLUX.1 dev	0.0733	0.0422	-0.0310	0.0049	-0.1255	0.0634	includes 0
DEAR	FLUX.1 dev	0.7636	0.7705	0.0069	0.0020	-0.0876	0.1013	includes 0
DeeCLIP	FLUX.1 dev	0.0052	0.0040	-0.0012	0.0004	-0.0956	0.0933	includes 0
DE-FAKE	FLUX.1 dev	0.5491	0.6828	0.1337	0.0045	0.0393	0.2282	> 0
DGS-Net	FLUX.1 dev	0.9302	0.9339	0.0036	0.0019	-0.0908	0.0981	includes 0
DIRE	FLUX.1 dev	-0.0002	-0.0006	-0.0004	0.0004	-0.0948	0.0941	includes 0
DRCT	FLUX.1 dev	0.0997	0.2597	0.1600	0.0068	0.0655	0.2544	> 0
Dual Data Alignment	FLUX.1 dev	0.6651	0.8396	0.1745	0.0061	0.0800	0.2690	> 0
FatFormer	FLUX.1 dev	0.0217	0.0305	0.0089	0.0044	-0.0856	0.1033	includes 0
FerretNet	FLUX.1 dev	0.9863	0.9609	-0.0254	0.0023	-0.1198	0.0691	includes 0
FIRE	FLUX.1 dev	0.0000	0.0000	0.0000	0.0000	-0.0945	0.0945	includes 0
Forensic-MoE	FLUX.1 dev	0.2795	0.3792	0.0997	0.0089	0.0052	0.1941	> 0
FSD	FLUX.1 dev	0.0023	-0.0024	-0.0047	0.0007	-0.0992	0.0897	includes 0
ForgeLens	FLUX.1 dev	0.6922	0.8344	0.1423	0.0055	0.0478	0.2367	> 0
FreqDetect	FLUX.1 dev	-0.2184	-0.2033	0.0151	0.0031	-0.0793	0.1096	includes 0
FreqNet	FLUX.1 dev	0.1074	0.4416	0.3342	0.0099	0.2398	0.4287	> 0
Fusing	FLUX.1 dev	0.0194	0.0177	-0.0017	0.0027	-0.0962	0.0927	includes 0
GAPL	FLUX.1 dev	0.8781	0.9429	0.0647	0.0043	-0.0297	0.1592	includes 0
IAPL	FLUX.1 dev	0.8246	0.8750	0.0504	0.0047	-0.0440	0.1449	includes 0
InfoFD	FLUX.1 dev	0.2803	0.3063	0.0261	0.0010	-0.0684	0.1205	includes 0
LaDeDa	FLUX.1 dev	0.2362	0.0375	-0.1986	0.0081	-0.2931	-0.1042	< 0
LASTED	FLUX.1 dev	0.0057	0.0093	0.0036	0.0016	-0.0908	0.0981	includes 0
LGrad	FLUX.1 dev	0.3192	0.4415	0.1224	0.0071	0.0279	0.2168	> 0
MDMF	FLUX.1 dev	-0.0000	-0.0000	0.0000	0.0000	-0.0944	0.0945	includes 0
MIRROR	FLUX.1 dev	0.7745	0.8744	0.0999	0.0043	0.0054	0.1943	> 0
MLEP	FLUX.1 dev	0.5259	0.5331	0.0072	0.0014	-0.0872	0.1017	includes 0
NPR	FLUX.1 dev	0.4992	0.5170	0.0178	0.0025	-0.0766	0.1123	includes 0
OmniAID	FLUX.1 dev	0.7590	0.7614	0.0024	0.0016	-0.0920	0.0969	includes 0
PatchCraft	FLUX.1 dev	0.8425	0.8648	0.0222	0.0056	-0.0722	0.1167	includes 0
PGC	FLUX.1 dev	0.0782	0.0870	0.0088	0.0028	-0.0856	0.1033	includes 0
PROBE	FLUX.1 dev	0.3920	0.4772	0.0852	0.0095	-0.0092	0.1797	includes 0
Raising the Bar	FLUX.1 dev	0.1560	0.3202	0.1643	0.0039	0.0698	0.2587	> 0
RealStats	FLUX.1 dev	-0.0503	0.0724	0.1227	0.0070	0.0282	0.2171	> 0
RIGID	FLUX.1 dev	0.0013	0.0035	0.0022	0.0002	-0.0922	0.0967	includes 0
RINE	FLUX.1 dev	-0.0030	-0.0033	-0.0003	0.0006	-0.0947	0.0942	includes 0
SAFE	FLUX.1 dev	0.8576	0.8752	0.0176	0.0022	-0.0768	0.1121	includes 0
SDAIE	FLUX.1 dev	0.9728	0.9769	0.0041	0.0009	-0.0903	0.0986	includes 0
SPAI	FLUX.1 dev	0.4225	0.5905	0.1679	0.0100	0.0735	0.2624	> 0
SSP	FLUX.1 dev	-0.7726	-0.9189	-0.1464	0.0059	-0.2408	-0.0519	< 0
Stay-Positive	FLUX.1 dev	0.4444	0.4069	-0.0375	0.0024	-0.1320	0.0569	includes 0
Synthbuster	FLUX.1 dev	0.0000	0.0000	0.0000	0.0000	-0.0945	0.0945	includes 0
Synthetic Laundering	FLUX.1 dev	0.6565	0.6771	0.0206	0.0019	-0.0739	0.1150	includes 0
UniversalFakeDetect	FLUX.1 dev	-0.0009	-0.0080	-0.0070	0.0014	-0.1015	0.0874	includes 0
WaRPAD	FLUX.1 dev	0.0005	0.0007	0.0002	0.0000	-0.0943	0.0946	includes 0
WePe	FLUX.1 dev	-0.0025	-0.0038	-0.0013	0.0001	-0.0958	0.0931	includes 0

FLUX.1 dev. Table 21 contains 17 resolved route contrasts, comprising 15 positive and two negative intervals. The route-contrast median is 0.0088, with an interquartile range of 0.0964. FreqNet gives the largest positive contrast at 0.3342, while LaDeDa gives the smallest contrast at -0.1986. MIRROR contributes a positive contrast of 0.0999. This two-sided detector-level spread coexists with a profile correlation of 0.977.

Table 22: Complete Qwen Image text to image route contrasts. Point estimates are exact frozen-roster summaries, SE is the descriptive source-cluster standard error, and intervals are A3/A4(i)-conditional.

Detector	Family	$m(0.8)$	ρ^{T2I}	ζ^{T2I}	SE	Lower	Upper	CI status
AEROBLADE	Qwen Image	0.0029	0.0061	0.0032	0.0000	-0.0912	0.0977	includes 0
AIDE	Qwen Image	0.6835	0.8064	0.1229	0.0041	0.0285	0.2174	> 0
AlignedForensics	Qwen Image	0.5309	0.6454	0.1145	0.0077	0.0201	0.2090	> 0
B-Free	Qwen Image	0.6370	0.6638	0.0268	0.0090	-0.0677	0.1213	includes 0
C2P-CLIP	Qwen Image	0.2848	0.2131	-0.0717	0.0059	-0.1662	0.0227	includes 0
CLIDE	Qwen Image	0.0000	0.0000	0.0000	0.0000	-0.0944	0.0945	includes 0
CLIPping	Qwen Image	0.0492	-0.0214	-0.0707	0.0028	-0.1651	0.0238	includes 0
CNNSpot	Qwen Image	-0.0014	-0.0028	-0.0013	0.0008	-0.0958	0.0931	includes 0
Community Forensics	Qwen Image	0.5818	0.9594	0.3776	0.0082	0.2832	0.4721	> 0
ConV	Qwen Image	0.0011	-0.0038	-0.0048	0.0001	-0.0993	0.0896	includes 0
Corvi detector	Qwen Image	0.9576	0.9931	0.0355	0.0032	-0.0590	0.1299	includes 0
CO-SPY	Qwen Image	0.7811	0.8644	0.0833	0.0043	-0.0111	0.1778	includes 0
D3	Qwen Image	0.6479	0.7646	0.1167	0.0075	0.0223	0.2112	> 0
D3QE	Qwen Image	0.1856	0.1266	-0.0589	0.0063	-0.1534	0.0355	includes 0
DEAR	Qwen Image	0.8598	0.8673	0.0075	0.0007	-0.0870	0.1019	includes 0
DeeCLIP	Qwen Image	0.0511	0.0415	-0.0096	0.0009	-0.1041	0.0848	includes 0
DE-FAKE	Qwen Image	0.5846	0.6974	0.1128	0.0040	0.0183	0.2072	> 0
DGS-Net	Qwen Image	0.9319	0.9392	0.0072	0.0013	-0.0872	0.1017	includes 0
DIRE	Qwen Image	-0.0011	-0.0015	-0.0005	0.0007	-0.0949	0.0940	includes 0
DRCT	Qwen Image	0.2467	0.3453	0.0986	0.0079	0.0041	0.1930	> 0
Dual Data Alignment	Qwen Image	0.8409	0.9277	0.0868	0.0039	-0.0077	0.1812	includes 0
FatFormer	Qwen Image	0.0726	0.0776	0.0050	0.0064	-0.0895	0.0994	includes 0
FerretNet	Qwen Image	0.9769	0.9896	0.0127	0.0017	-0.0817	0.1072	includes 0
FIRE	Qwen Image	-0.0000	0.0000	0.0000	0.0000	-0.0944	0.0945	includes 0
Forensic-MoE	Qwen Image	0.6274	0.5886	-0.0388	0.0090	-0.1333	0.0556	includes 0
FSD	Qwen Image	0.0054	0.0045	-0.0009	0.0002	-0.0954	0.0935	includes 0
ForgeLens	Qwen Image	0.8586	0.8533	-0.0054	0.0024	-0.0998	0.0891	includes 0
FreqDetect	Qwen Image	-0.1820	-0.0924	0.0896	0.0056	-0.0049	0.1840	includes 0
FreqNet	Qwen Image	0.5368	0.7700	0.2332	0.0093	0.1387	0.3276	> 0
Fusing	Qwen Image	0.1579	0.2430	0.0851	0.0079	-0.0094	0.1795	includes 0
GAPL	Qwen Image	0.8450	0.9592	0.1142	0.0050	0.0197	0.2086	> 0
IAPL	Qwen Image	0.8238	0.8765	0.0526	0.0046	-0.0418	0.1471	includes 0
InfoFD	Qwen Image	0.3062	0.2879	-0.0183	0.0011	-0.1127	0.0762	includes 0
LaDeDa	Qwen Image	0.4577	0.5105	0.0528	0.0099	-0.0416	0.1473	includes 0
LASTED	Qwen Image	0.1764	0.0088	-0.1676	0.0055	-0.2621	-0.0732	< 0
LGrad	Qwen Image	0.4452	0.4719	0.0267	0.0053	-0.0677	0.1212	includes 0
MDMF	Qwen Image	0.0001	0.0000	-0.0001	0.0000	-0.0945	0.0944	includes 0
MIRROR	Qwen Image	0.8539	0.8896	0.0357	0.0027	-0.0588	0.1301	includes 0
MLEP	Qwen Image	0.5258	0.5346	0.0087	0.0013	-0.0857	0.1032	includes 0
NPR	Qwen Image	0.5048	0.5182	0.0134	0.0020	-0.0811	0.1078	includes 0
OmniAID	Qwen Image	0.6564	0.7133	0.0570	0.0032	-0.0375	0.1514	includes 0
PatchCraft	Qwen Image	0.9521	0.9740	0.0219	0.0022	-0.0725	0.1164	includes 0
PGC	Qwen Image	0.3730	0.2426	-0.1304	0.0046	-0.2248	-0.0359	< 0
PROBE	Qwen Image	0.7771	0.9629	0.1857	0.0066	0.0913	0.2802	> 0
Raising the Bar	Qwen Image	0.2791	0.3924	0.1134	0.0043	0.0189	0.2078	> 0
RealStats	Qwen Image	0.0021	0.0592	0.0572	0.0068	-0.0373	0.1516	includes 0
RIGID	Qwen Image	0.0003	0.0021	0.0018	0.0001	-0.0927	0.0962	includes 0
RINE	Qwen Image	0.0225	-0.0028	-0.0253	0.0023	-0.1198	0.0692	includes 0
SAFE	Qwen Image	0.7145	0.7485	0.0340	0.0056	-0.0605	0.1284	includes 0
SDAIE	Qwen Image	0.9661	0.9796	0.0135	0.0015	-0.0810	0.1079	includes 0
SPAI	Qwen Image	0.6769	0.6589	-0.0180	0.0085	-0.1125	0.0764	includes 0
SSP	Qwen Image	-0.5202	-0.8686	-0.3484	0.0091	-0.4429	-0.2540	< 0
Stay-Positive	Qwen Image	0.6995	0.7052	0.0057	0.0009	-0.0887	0.1002	includes 0
Synthbuster	Qwen Image	-0.0001	-0.0001	0.0000	0.0000	-0.0944	0.0945	includes 0
Synthetic Laundering	Qwen Image	0.6977	0.7004	0.0027	0.0023	-0.0918	0.0971	includes 0
UniversalFakeDetect	Qwen Image	0.0608	-0.0012	-0.0620	0.0032	-0.1565	0.0324	includes 0
WaRPAD	Qwen Image	0.0012	0.0008	-0.0004	0.0000	-0.0949	0.0940	includes 0
WePe	Qwen Image	-0.0020	-0.0047	-0.0026	0.0001	-0.0971	0.0918	includes 0

Qwen Image. Table 22 contains 13 resolved route contrasts, with ten positive and three negative intervals. The median is 0.0073 and the interquartile range is 0.0594. Community Forensics gives the largest positive contrast at 0.3776, whereas SSP gives the smallest contrast at -0.3484. These extrema form the widest detector-level span among the four route comparisons, while the complete profile correlation remains 0.978.

Table 23: Complete Kandinsky 2.2 text to image route contrasts. Point estimates are exact frozen-roster summaries, SE is the descriptive source-cluster standard error, and intervals are A3/A4(i)-conditional.

Detector	Family	$m(0.8)$	ρ^{T2I}	ζ^{T2I}	SE	Lower	Upper	CI status
AEROBLADE	Kandinsky 2.2	0.0091	0.0091	0.0000	0.0000	-0.0944	0.0945	includes 0
AIDE	Kandinsky 2.2	0.5509	0.5255	-0.0254	0.0047	-0.1198	0.0691	includes 0
AlignedForensics	Kandinsky 2.2	0.8177	0.8192	0.0015	0.0030	-0.0930	0.0959	includes 0
B-Free	Kandinsky 2.2	0.9878	0.9874	-0.0005	0.0006	-0.0949	0.0940	includes 0
C2P-CLIP	Kandinsky 2.2	0.2865	0.2898	0.0034	0.0033	-0.0911	0.0978	includes 0
CLIDE	Kandinsky 2.2	0.0000	0.0000	0.0000	0.0000	-0.0945	0.0945	includes 0
CLIPping	Kandinsky 2.2	0.1486	0.1536	0.0050	0.0023	-0.0895	0.0994	includes 0
CNNSpot	Kandinsky 2.2	0.0018	0.0033	0.0015	0.0012	-0.0929	0.0960	includes 0
Community Forensics	Kandinsky 2.2	0.9895	0.9866	-0.0028	0.0013	-0.0973	0.0916	includes 0
ConV	Kandinsky 2.2	0.0058	0.0064	0.0006	0.0002	-0.0939	0.0950	includes 0
Corvi detector	Kandinsky 2.2	0.9291	0.9285	-0.0006	0.0020	-0.0951	0.0938	includes 0
CO-SPY	Kandinsky 2.2	0.8036	0.8039	0.0002	0.0029	-0.0942	0.0947	includes 0
D3	Kandinsky 2.2	0.8981	0.8986	0.0006	0.0025	-0.0939	0.0950	includes 0
D3QE	Kandinsky 2.2	0.2833	0.2907	0.0074	0.0046	-0.0870	0.1019	includes 0
DEAR	Kandinsky 2.2	0.8167	0.8125	-0.0042	0.0006	-0.0986	0.0903	includes 0
DeeCLIP	Kandinsky 2.2	0.0536	0.0541	0.0004	0.0006	-0.0940	0.0949	includes 0
DE-FAKE	Kandinsky 2.2	0.6831	0.6807	-0.0024	0.0013	-0.0968	0.0921	includes 0
DGS-Net	Kandinsky 2.2	0.9149	0.9099	-0.0050	0.0025	-0.0995	0.0894	includes 0
DIRE	Kandinsky 2.2	-0.0036	-0.0043	-0.0008	0.0011	-0.0952	0.0937	includes 0
DRCT	Kandinsky 2.2	0.5803	0.5711	-0.0092	0.0044	-0.1036	0.0853	includes 0
Dual Data Alignment	Kandinsky 2.2	0.9300	0.9281	-0.0019	0.0007	-0.0963	0.0926	includes 0
FatFormer	Kandinsky 2.2	0.2912	0.2979	0.0067	0.0058	-0.0877	0.1012	includes 0
FerretNet	Kandinsky 2.2	0.9532	0.9486	-0.0046	0.0018	-0.0990	0.0899	includes 0
FIRE	Kandinsky 2.2	0.0000	0.0000	-0.0000	0.0000	-0.0945	0.0945	includes 0
Forensic-MoE	Kandinsky 2.2	0.9704	0.9729	0.0025	0.0020	-0.0919	0.0970	includes 0
FSD	Kandinsky 2.2	-0.0036	-0.0034	0.0002	0.0008	-0.0942	0.0947	includes 0
ForgeLens	Kandinsky 2.2	0.9102	0.9101	-0.0001	0.0002	-0.0945	0.0944	includes 0
FreqDetect	Kandinsky 2.2	0.0883	0.0809	-0.0075	0.0061	-0.1019	0.0870	includes 0
FreqNet	Kandinsky 2.2	0.4046	0.3942	-0.0104	0.0058	-0.1048	0.0841	includes 0
Fusing	Kandinsky 2.2	0.1590	0.1676	0.0086	0.0052	-0.0859	0.1030	includes 0
GAPL	Kandinsky 2.2	0.9652	0.9643	-0.0009	0.0007	-0.0954	0.0935	includes 0
IAPL	Kandinsky 2.2	0.9012	0.9012	0.0000	0.0000	-0.0944	0.0945	includes 0
InfoFD	Kandinsky 2.2	0.3623	0.3615	-0.0008	0.0005	-0.0953	0.0936	includes 0
LaDeDa	Kandinsky 2.2	0.2721	0.2472	-0.0249	0.0064	-0.1193	0.0696	includes 0
LASTED	Kandinsky 2.2	0.0347	0.0371	0.0024	0.0020	-0.0921	0.0968	includes 0
LGrad	Kandinsky 2.2	0.4020	0.3928	-0.0093	0.0049	-0.1037	0.0852	includes 0
MDMF	Kandinsky 2.2	0.0001	0.0001	-0.0000	0.0000	-0.0945	0.0945	includes 0
MIRROR	Kandinsky 2.2	0.8924	0.8921	-0.0003	0.0003	-0.0948	0.0941	includes 0
MLEP	Kandinsky 2.2	0.5073	0.4941	-0.0133	0.0026	-0.1077	0.0812	includes 0
NPR	Kandinsky 2.2	0.4781	0.4740	-0.0041	0.0028	-0.0985	0.0904	includes 0
OmniAID	Kandinsky 2.2	0.6712	0.6706	-0.0007	0.0016	-0.0951	0.0938	includes 0
PatchCraft	Kandinsky 2.2	0.8372	0.8449	0.0077	0.0037	-0.0868	0.1021	includes 0
PGC	Kandinsky 2.2	0.5634	0.5567	-0.0067	0.0017	-0.1011	0.0878	includes 0
PROBE	Kandinsky 2.2	0.9903	0.9885	-0.0018	0.0007	-0.0962	0.0927	includes 0
Raising the Bar	Kandinsky 2.2	0.4434	0.4453	0.0019	0.0021	-0.0925	0.0964	includes 0
RealStats	Kandinsky 2.2	0.1095	0.1134	0.0040	0.0040	-0.0905	0.0984	includes 0
RIGID	Kandinsky 2.2	0.0078	0.0080	0.0002	0.0001	-0.0942	0.0947	includes 0
RINE	Kandinsky 2.2	0.0922	0.1004	0.0082	0.0033	-0.0863	0.1026	includes 0
SAFE	Kandinsky 2.2	0.8569	0.8552	-0.0016	0.0021	-0.0961	0.0928	includes 0
SDAIE	Kandinsky 2.2	0.9623	0.9614	-0.0010	0.0014	-0.0954	0.0935	includes 0
SPAI	Kandinsky 2.2	0.2868	0.2713	-0.0154	0.0060	-0.1099	0.0790	includes 0
SSP	Kandinsky 2.2	-0.6530	-0.6052	0.0478	0.0063	-0.0467	0.1422	includes 0
Stay-Positive	Kandinsky 2.2	0.6249	0.6184	-0.0065	0.0009	-0.1009	0.0880	includes 0
Synthbuster	Kandinsky 2.2	0.0000	0.0000	0.0000	0.0000	-0.0945	0.0945	includes 0
Synthetic Laundering	Kandinsky 2.2	0.6723	0.6686	-0.0036	0.0011	-0.0981	0.0908	includes 0
UniversalFakeDetect	Kandinsky 2.2	0.2210	0.2232	0.0022	0.0047	-0.0923	0.0966	includes 0
WaRPAD	Kandinsky 2.2	0.0018	0.0018	0.0000	0.0000	-0.0944	0.0945	includes 0
WePe	Kandinsky 2.2	0.0009	0.0010	0.0000	0.0001	-0.0944	0.0945	includes 0

Kandinsky 2.2. Table 23 has a route-contrast median of approximately zero and an interquartile range of 0.0047. Its detector-level contrasts range from -0.0254 for AIDE to 0.0478 for SSP, and all 58 simultaneous intervals include zero. The source-conditioned and text-only detector profiles therefore align almost exactly in this registered implementation, with a profile correlation of 1.000.

Table 24: Complete Kolors text to image route contrasts. Point estimates are exact frozen-roster summaries, SE is the descriptive source-cluster standard error, and intervals are A3/A4(i)-conditional.

Detector	Family	$m(0.8)$	ρ^{T2I}	ζ^{T2I}	SE	Lower	Upper	CI status
AEROBLADE	Kolors	0.0059	0.0063	0.0004	0.0000	-0.0941	0.0948	includes 0
AIDE	Kolors	0.7106	0.6767	-0.0340	0.0050	-0.1284	0.0605	includes 0
AlignedForensics	Kolors	0.7373	0.7904	0.0530	0.0061	-0.0414	0.1475	includes 0
B-Free	Kolors	0.9906	0.9907	0.0001	0.0000	-0.0944	0.0945	includes 0
C2P-CLIP	Kolors	0.4940	0.4192	-0.0747	0.0074	-0.1692	0.0197	includes 0
CLIDE	Kolors	0.0001	0.0001	-0.0000	0.0000	-0.0945	0.0944	includes 0
CLIPping	Kolors	0.0620	0.0081	-0.0538	0.0030	-0.1483	0.0406	includes 0
CNNSpot	Kolors	-0.0029	-0.0028	0.0001	0.0006	-0.0944	0.0945	includes 0
Community Forensics	Kolors	0.9987	0.9985	-0.0002	0.0002	-0.0947	0.0942	includes 0
ConV	Kolors	0.0079	0.0045	-0.0034	0.0002	-0.0978	0.0911	includes 0
Corvi detector	Kolors	0.8628	0.9024	0.0396	0.0040	-0.0549	0.1340	includes 0
CO-SPY	Kolors	0.6773	0.6699	-0.0074	0.0073	-0.1018	0.0871	includes 0
D3	Kolors	0.8427	0.8201	-0.0226	0.0053	-0.1170	0.0719	includes 0
D3QE	Kolors	0.3607	0.2881	-0.0726	0.0068	-0.1671	0.0218	includes 0
DEAR	Kolors	0.7898	0.7911	0.0012	0.0012	-0.0932	0.0957	includes 0
DeeCLIP	Kolors	0.0563	0.0511	-0.0051	0.0008	-0.0996	0.0893	includes 0
DE-FAKE	Kolors	0.6827	0.6913	0.0086	0.0013	-0.0858	0.1031	includes 0
DGS-Net	Kolors	0.9359	0.9298	-0.0061	0.0016	-0.1006	0.0883	includes 0
DIRE	Kolors	-0.0017	-0.0031	-0.0013	0.0009	-0.0958	0.0931	includes 0
DRCT	Kolors	0.6465	0.6856	0.0392	0.0062	-0.0553	0.1336	includes 0
Dual Data Alignment	Kolors	0.9380	0.9375	-0.0005	0.0001	-0.0950	0.0939	includes 0
FatFormer	Kolors	0.6502	0.5555	-0.0947	0.0097	-0.1891	-0.0002	< 0
FerretNet	Kolors	0.9835	0.9718	-0.0117	0.0014	-0.1062	0.0828	includes 0
FIRE	Kolors	-0.0000	-0.0000	-0.0000	0.0000	-0.0945	0.0945	includes 0
Forensic-MoE	Kolors	0.5068	0.4682	-0.0386	0.0092	-0.1330	0.0559	includes 0
FSD	Kolors	-0.0142	-0.0183	-0.0041	0.0013	-0.0985	0.0904	includes 0
ForgeLens	Kolors	0.8202	0.8015	-0.0187	0.0038	-0.1131	0.0758	includes 0
FreqDetect	Kolors	-0.1612	-0.1528	0.0084	0.0044	-0.0861	0.1028	includes 0
FreqNet	Kolors	0.0574	0.1051	0.0477	0.0076	-0.0468	0.1422	includes 0
Fusing	Kolors	-0.0025	0.0039	0.0064	0.0015	-0.0880	0.1009	includes 0
GAPL	Kolors	0.9665	0.9662	-0.0003	0.0004	-0.0948	0.0941	includes 0
IAPL	Kolors	0.8652	0.8548	-0.0104	0.0039	-0.1048	0.0841	includes 0
InfoFD	Kolors	0.2805	0.2811	0.0006	0.0010	-0.0939	0.0950	includes 0
LaDeDa	Kolors	0.0018	0.0077	0.0059	0.0058	-0.0886	0.1003	includes 0
LASTED	Kolors	0.0616	0.0878	0.0262	0.0039	-0.0682	0.1207	includes 0
LGrad	Kolors	0.2868	0.2935	0.0067	0.0085	-0.0878	0.1011	includes 0
MDMF	Kolors	0.0018	0.0008	-0.0010	0.0001	-0.0954	0.0935	includes 0
MIRROR	Kolors	0.8931	0.8930	-0.0001	0.0001	-0.0945	0.0944	includes 0
MLEP	Kolors	0.5346	0.5311	-0.0034	0.0009	-0.0979	0.0910	includes 0
NPR	Kolors	0.4883	0.4923	0.0039	0.0034	-0.0905	0.0984	includes 0
OmniAID	Kolors	0.7893	0.7928	0.0035	0.0012	-0.0910	0.0979	includes 0
PatchCraft	Kolors	0.7638	0.7850	0.0212	0.0068	-0.0733	0.1156	includes 0
PGC	Kolors	0.6473	0.6196	-0.0277	0.0017	-0.1221	0.0668	includes 0
PROBE	Kolors	0.9943	0.9935	-0.0008	0.0002	-0.0952	0.0937	includes 0
Raising the Bar	Kolors	0.5199	0.5351	0.0152	0.0036	-0.0792	0.1097	includes 0
RealStats	Kolors	0.0573	-0.0111	-0.0684	0.0061	-0.1629	0.0260	includes 0
RIGID	Kolors	0.0076	0.0055	-0.0021	0.0002	-0.0965	0.0924	includes 0
RINE	Kolors	0.3043	0.1457	-0.1587	0.0079	-0.2531	-0.0642	< 0
SAFE	Kolors	0.8571	0.8604	0.0033	0.0025	-0.0911	0.0978	includes 0
SDAIE	Kolors	0.9366	0.9491	0.0125	0.0027	-0.0820	0.1069	includes 0
SPAI	Kolors	0.7246	0.7109	-0.0137	0.0071	-0.1082	0.0807	includes 0
SSP	Kolors	-0.7621	-0.7909	-0.0288	0.0071	-0.1233	0.0656	includes 0
Stay-Positive	Kolors	0.5807	0.5835	0.0028	0.0021	-0.0916	0.0973	includes 0
Synthbuster	Kolors	0.0000	0.0000	0.0000	0.0000	-0.0945	0.0945	includes 0
Synthetic Laundering	Kolors	0.7006	0.7109	0.0103	0.0016	-0.0842	0.1047	includes 0
UniversalFakeDetect	Kolors	0.1613	0.0814	-0.0799	0.0052	-0.1743	0.0146	includes 0
WaRPAD	Kolors	0.0018	0.0017	-0.0001	0.0000	-0.0946	0.0944	includes 0
WePe	Kolors	0.0024	-0.0002	-0.0025	0.0001	-0.0970	0.0919	includes 0

Kolors. Table 24 has a median of -0.0003 and an interquartile range of 0.0152. Two route contrasts are resolved below zero: -0.1587 for RINE and -0.0947 for FatFormer. The largest positive point contrast is 0.0530 for AlignedForensics, whose interval includes zero. The profile correlation is 0.996. The resulting pattern combines broad route agreement with two detector-specific negative changes.