DO FORENSIC VERDICTS FOLLOW OUTPUT REGIONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Explainable forensic VLMs pair manipulation verdicts with localized evidence, but localization overlap alone cannot establish whether the verdict actually depends on the cited region. Existing deletion and inpainting tests are also problematic because the intervention can itself introduce forensic cues. We introduce *paste-back auditing*: when a paired manipulation retains its authentic source, we restore authentic pixels only inside the model's output region and measure the resulting verdict change (necessity) alongside localization. Paste-back introduces no synthesized content and is exactly the identity on authentic images, whereas two classical inpainters flip 31.3% of authentic examples for served chat VLMs. In a positive control, a faithful oracle and a correct-but-non-causal oracle have identical IoU, yet necessity is 0.99 versus 0.00, showing that localization and intervention response capture distinct properties. Controlled Qwen2-VL-2B experiments reproduce this dissociation under a JPEG-75 cue. We further audit FakeShield and FFAA as released-system case studies. The audits reveal that output regions can be only partially aligned with the measured verdict response, while narration-grounded interventions can produce stronger effects when the narration explicitly identifies the manipulated region. Paste-back therefore provides a synthesis-free test of verdict–output-region alignment wherever the authentic source is retained, while keeping localization correctness and intervention dependence as separate quantities.

1 INTRODUCTION

Forensic vision–language models increasingly provide not only a manipulation verdict but also localized or narrated evidence explaining *where* and *why* an image appears manipulated. Systems such as FakeShield, FFAA, ForgeryGPT, and SIDA (Xu et al., 2025b; Huang et al., 2024; Zhang et al., 2024a; Huang et al., 2025a), together with recent successors (Wen et al., 2025; Kang et al., 2025; Sun et al., 2025; Zhou et al., 2025; Huang et al., 2025b; Jiang et al., 2026; Ji et al., 2026; Tan et al., 2026), produce explanations such as "the ear boundary is inconsistent." Yet current benchmarks largely evaluate whether this evidence *looks correct*, using region overlap, similarity to reference explanations, or LLM-based judgments (Papineni et al., 2002; Lin, 2004; Vedantam et al., 2015; Zhang et al., 2020; Lin et al., 2025; Narang et al., 2026; Yang et al., 2025; Kuckreja et al., 2026).

These evaluations do not answer a different question: *does the verdict actually depend on the region presented as evidence?* A detector may correctly localize a manipulation while making its decision from another cue, such as a canvas-wide generation or compression artifact. Its overlap score can then be identical to that of a detector whose verdict truly responds to the cited pixels. This is the familiar distinction between being right and being right for the right reason (Ross et al., 2017; Turpin et al., 2023), specialized here to forensic VLMs.

Testing this dependence requires intervening on the claimed evidence, but standard image interventions introduce their own confound. Masking or deleting a region creates an artificial hole, while inpainting introduces synthesized pixels that a forensic detector may itself react to (Chang et al., 2019; Goyal et al., 2019; Kaushik et al., 2020). **Our key observation is that paired local manipulations already provide the required counterfactual pixels.** When the authentic pre-edit source is retained, we can restore authentic pixels only inside the model's output region and re-evaluate the verdict. We call this intervention *paste-back auditing* (Figure 1). Paste-back introduces no synthesized content and is exactly the identity on authentic images, whereas the classical inpainters we

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.

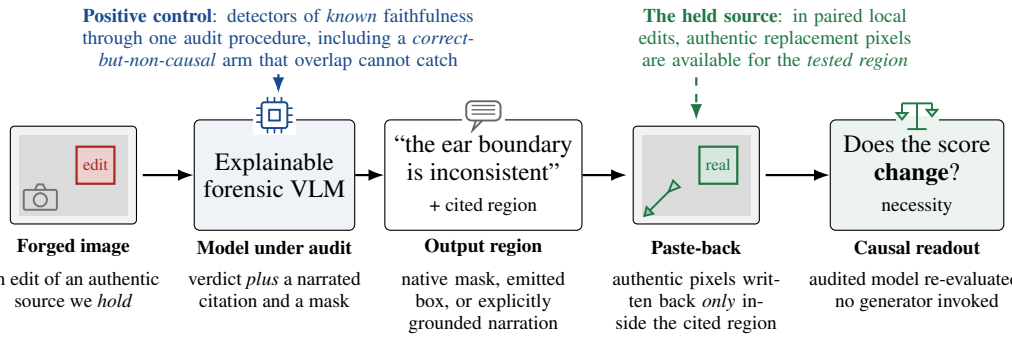


Figure 1: Paste-back auditing on paired local edits: authentic source pixels replace the output region, and the fixed verdict score is measured again.

evaluate flip 31.3% of authentic examples for served chat VLMs. Partial restoration can still induce boundary, context, and cue-interaction effects, which we measure rather than assume away.

Paste-back separates two quantities that are often conflated. *Localization* asks whether the output region overlaps the true manipulation, whereas *necessity* asks how much the specified verdict score changes when that region is restored to authentic content. In our positive control, a faithful oracle and a correct-but-non-causal oracle emit the same ground-truth region and therefore have identical IoU, while their necessities are 0.99 and 0.00, respectively. Conversely, a detector may depend strongly on a cited region while localizing only part of the manipulation. Localization and intervention response therefore capture complementary properties and should be reported separately.

We apply this framework to released forensic systems and controlled learned models. For FakeShield, we evaluate both a post-hoc grounding of its explanation and its own MFLM localization mask; because the latter is not established to depend on the generated narration, we treat it as an *output region*. For FFAA, the narration explicitly names face parts, allowing a direct comparison against a shuffled-narration control. We also train Qwen2-VL-2B detectors on controlled local and canvas-wide cues and reproduce the localization–intervention dissociation under a JPEG-75 cue. Together, these experiments show that plausible localization need not imply strong verdict dependence on the localized evidence.

Contributions. Our contributions are:

- We introduce **paste-back auditing**, a synthesis-free intervention for measuring verdict–output-region alignment when an authentic source is retained.
- We develop a **positive-control validation protocol** showing that localization overlap and intervention response capture distinct properties.
- We apply the framework to **released forensic VLMs and controlled learned VLMs**, revealing cases where plausible localization does not imply strong verdict dependence on the localized evidence.

2 Methodology

2.1 Problem Setup and Paste-Back Intervention

Let f be a forensic VLM with a fixed deterministic manipulation score $p(x)$. For an input image x, the system produces an output region $R(x)$, which may be a native mask, an emitted box, or a region obtained by mapping its explanation $e(x)$ through a stated grounder g. Let $M^\star$ denote the ground-truth manipulation support and x_{src} the corresponding pre-edit authentic source.

Our goal is to test whether the specified verdict responds to the region the system presents as evidence. When the authentic source is retained, the counterfactual "what if this region were authentic?" can be constructed without synthesis. For any region $R \subseteq \Omega$, we define the *paste-back*

intervention as

$$A_R^{\mathrm{pb}}(x)[u] = \begin{cases} x_{\mathrm{src}}[u], & u \in R, \\ x[u], & u \notin R. \end{cases}$$

Thus, only pixels inside R are replaced, using observed pixels from the authentic source rather than generated or imputed content.

Proposition 1 (Exact identity on authentic content)**.** *If $x = x_{\mathrm{src}}$, then $A_R^{\mathrm{pb}}(x) = x$ for every $R \subseteq \Omega$. Consequently, for any fixed deterministic detector f, the authentic-image false-flip rate is exactly* 0 *and*

$$p(x) - p(A_R^{\mathrm{pb}}(x)) = 0.$$

The proposition gives paste-back a simple but useful guarantee: applying the operator to authentic content cannot introduce a new synthesis artifact, because both branches read pixels from the same image. This distinguishes paste-back from deletion and generative inpainting, whose interventions can themselves alter the evidence seen by a forensic detector.

2.2 Measuring Verdict–Region Alignment

For a system output region $R(x)$, we define its *necessity* as

$$\mathrm{Nec}(x) = p(x) - p(A_{R(x)}^{\mathrm{pb}}(x)).$$

Necessity measures how much the specified manipulation score changes when the claimed region is restored to authentic content. A larger positive value indicates a stronger verdict response to restoration of that region.

A detector may, however, respond to restoration in many parts of an image. We therefore compare the cited-region intervention with a control region R_{ctrl} that is disjoint from $R(x)$, area matched, and matched on the detector's own occlusion sensitivity:

$$\mathrm{Spec}(x) = p(x) - p(A_{R_{\mathrm{ctrl}}}^{\mathrm{pb}}(x)), \qquad \Delta\mathrm{NS}(x) = \mathrm{Nec}(x) - \mathrm{Spec}(x).$$

Algorithm 1 gives the complete audit procedure.

A congruent disjoint control does not always fit, particularly when the output region occupies a large fraction of the image. Restricting evaluation to rows where such a control exists would systematically remove precisely these large-region cases. In our served-model evaluation, such a restriction would discard $1 - 0.83$ of the rows. When necessary, we therefore use an area-exact fragmented control and quantify the effect of this substitution in Apx. L.2. The control is constrained to be disjoint from $R(x)$, but not necessarily from $M^\star$; consequently, Spec is not guaranteed to be zero or nonnegative.

We report intervention response alongside localization:

$$L_{\mathrm{gt}} = \mathrm{IoU}(R(x), M^\star).$$

The two quantities answer different questions. L_{gt} asks whether the system points to the manipulated content, whereas Nec asks whether the specified verdict responds when that content is restored. We therefore treat localization and intervention response as complementary measurements rather than combining them into a single primary score.

Score scale. Probability differences can become compressed when a detector operates near sigmoid saturation. We therefore also report the same intervention on the log-odds scale when useful. This is only a reparameterization of the score, not a stronger estimator; the corresponding sensitivity analysis is reported in Apx. H.3.1.

2.3 Guarantees and Limitations

Proposition 1 guarantees identity on an authentic input, but it does *not* imply that partial restoration of a manipulated image is distribution preserving. Patch boundaries, interactions between restored and surrounding pixels, or redundant manipulation cues may still affect the detector. We therefore distinguish the guarantee of the operator from empirical assumptions about the corpus and model.

Source locality. Whether restoring the full manipulation support recovers the source image depends on how the manipulated image was constructed. On our certified splice/copy-move corpus, source and manipulated images are bit-identical outside $M^\star$, so

$$A^{\mathrm{pb}}_{M^\star}(x) = x_{\mathrm{src}} \qquad \text{and} \qquad \mathrm{Nec}(M^\star) = p(x) - p(x_{\mathrm{src}}).$$

MagicBrush does not satisfy this condition because the manipulated image is a whole-canvas DALL-E-2 re-render of the source photograph. We therefore measure its outside-mask residual explicitly rather than assuming exact restoration (Apx. E.13). We likewise evaluate hard, feathered, and re-compressed constructions when boundary or compression effects may matter.

What paste-back measures. Paste-back measures *evidential necessity*: whether replacing the claimed manipulated content with its authentic counterpart changes the specified verdict. It is not a general test of computational dependence. For example, a model may internally use pixels in an authentic region, but pasting authentic pixels back into that region leaves the image unchanged and therefore cannot reveal that dependence. Claims from paste-back are consequently about verdict–region alignment under the stated intervention, not a complete description of the model's internal computation.

3 Experiments

3.1 Experimental Setup

We organize the evaluation around four questions: **(RQ1)** whether paste-back separates localization from verdict response under known dependence structure; **(RQ2)** whether released forensic VLMs' verdicts respond to their reported regions; **(RQ3)** whether the same dissociation can emerge in trained VLMs; and **(RQ4)** how sensitive these conclusions are to the intervention, readout, and corpus construction.

Datasets. We use MagicBrush for source-retained generative edits, a certified splice/copy-move corpus whose source and manipulated images are bit-identical outside the manipulation mask, and a source-retained face-part replacement corpus for FFAA. Controlled learned-model experiments use variants of the certified local corpus augmented with either a synthetic canvas-wide cue or whole-image JPEG-75 re-encoding. Dataset construction and certification are detailed in the appendix.

Systems. We evaluate two released forensic systems, FakeShield and FFAA, together with controlled oracle detectors and Qwen2-VL-2B models trained under prescribed local and global cue conditions. Served general-purpose VLMs are used only as paired contrasts and intervention sanity checks, not as the target population.

Regions and readouts. Every result states both the verdict readout and the source of the intervened region. Regions may be the ground-truth manipulation, a system-native mask, an emitted box, or a region obtained from narration by a stated mapping. We do not interpret a native region as narration-conditioned unless dependence on the narration is established.

Metrics. We report localization using $L_{\mathrm{gt}} = \mathrm{IoU}(R(x), M^\star)$ and intervention response using necessity Nec and, where appropriate, the control-adjusted quantity $\Delta\mathrm{NS}$. Normalized effects are used only when a meaningful full-restoration reference is available and are not interpreted as fractions of internal evidence.

Intervention sanity check. We first verify that the choice of intervention can itself affect the audit. Unlike paste-back, unconstrained inpainting has no authentic-image identity guarantee. Under our classical Laplace fill, sandbox necessity is -0.36 before calibration and +0.61 after calibration, against a paste-back reference of +0.34. On served chat VLMs, inpainting can also flip authentic inputs without removing any manipulation evidence. We therefore use paste-back as the primary intervention throughout; full operator comparisons and false-flip measurements are reported in Apx. G.

Uncertainty and tests. We report 95% bootstrap intervals and paired tests when comparisons are defined on the same images. Primary conclusions are based on prespecified contrasts; estimator, score-scale, and multiple-testing sensitivities are reported in the Apx. H.3.1.

3.2 POSITIVE-CONTROL VALIDATION

A low intervention response is only meaningful if the audit can recover models with known dependence structure. We therefore evaluate four oracle detectors on the same MagicBrush triples: faithful, decoy, half-evidence, and global-cue arms. The faithful oracle detects perfectly from the edited region, so restoring that region removes its evidence while restoring a disjoint region does not.

Table 1 summarizes the oracle results, and Figure 2 highlights the separation between localization and intervention response, including the fixed-localization $\texttt{mixed_cue}(\beta)$ sweep.

The key comparison is the *global-cue* oracle. It reports the same ground-truth box as the faithful oracle but bases its verdict on a canvas-wide generator fingerprint outside that box. The two therefore have identical localization and matched perfect detection (AUC 1.00), while differing in whether the cited region supports the verdict. This construction is also empirically motivated: a served chat VLM responds to the same canvas-wide re-encoding signal on this corpus (Apx. F).

The two contrasts show why both measurements are needed. Faithful and global-cue oracles have identical IoU ($\Delta\mathrm{IoU} = 0.00$[0.00, 0.00]) but necessity 0.99 versus 0.00, so localization alone cannot distinguish verdict dependence on the cited region. Conversely, the half-evidence oracle has necessity comparable to the faithful arm but lower IoU, showing that necessity does not measure localization completeness.

In $\texttt{mixed_cue}(\beta)$, localization is fixed (IoU 0.75) while increasing β shifts the verdict from global to local evidence, and necessity correspondingly traverses 0.00, 0.08, 0.52, 0.93, 0.99 (Figure 2b). We therefore report localization and intervention response separately throughout.

3.3 RELEASED-SYSTEM AUDITS

We apply paste-back to two released forensic VLMs, FakeShield (Xu et al., 2025b) and FFAA (Huang et al., 2024). Both are evaluated with fixed readouts and held-source image pairs after verifying the released checkpoints and preprocessing configurations (Appendices E and E′). Neither system emits a usable manipulation box, so the mapping from model output to intervened region is itself part of the audit. The two systems are *case studies, not a ranking*: they differ in corpus, readout, region instrument, and eligible cohort. Table 2 summarizes the principal intervention effects for both systems; the following analyses explain how each region and readout should be interpreted.

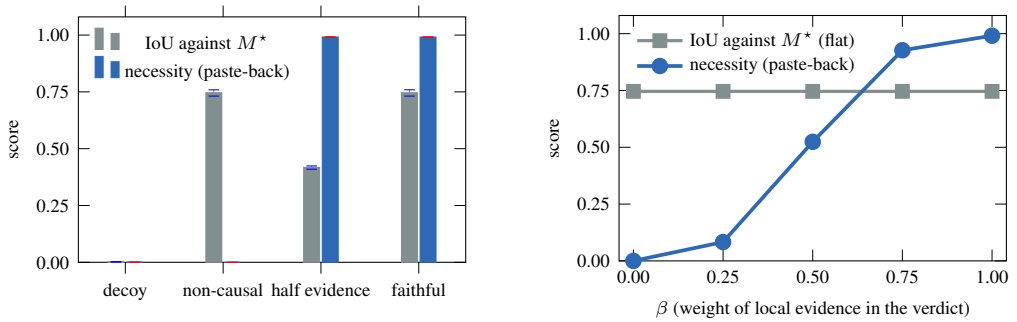
3.3.1 FAKESHIELD

Detection and a grounder-free anchor. FakeShield reaches AUC 0.59 $[0.52, 0.66]$ on MagicBrush and 0.981 $[0.955, 0.999]$ on the certified splice/copy-move corpus, indicating strong corpus dependence. On the certified corpus, restoring the true mask gives a grounder-free reference with $\Delta\mathrm{NS} = 0.29$[0.25, 0.32]. Because the source and manipulated images are identical outside $M^\star$, this restoration recovers the full detection margin and the matched-control effect is zero. This establishes sensitivity to the manipulated pixels before evaluating whether the reported region identifies them.

Post-hoc grounding. FakeShield emits no box, so we first ground its text explanation with a fixed post-hoc grounder. The effect is unresolved on MagicBrush but positive on the certified corpus ($\Delta\mathrm{NS} = +0.02$[+0.01, +0.03] at $n = 150$), although the mean is tail-sensitive and the median interval includes zero (Apx. I). Because the grounder itself performs near its random-location floor, a weak grounded effect may reflect either weak verdict–region alignment or poor grounding.

FakeShield's own output mask. FakeShield also provides MFLM, its native GLaMM/SAM localizer, which reaches IoU 0.496 $[0.393, 0.601]$ versus 0.13 $[0.09, 0.17]$ for the post-hoc grounder. Because shuffled narratives yield comparable masks (Apx. E.6), we treat MFLM as a system output region, not a narration-conditioned explanation.

On 49 eligible rows, restoring the MFLM mask gives normalized necessity 0.22 $[0.12, 0.33]$, versus 0.09 $[0.04, 0.16]$ through the post-hoc grounder (Figure 3a). This is not an "evidence share": the same mask has normalized sufficiency 0.89 $[0.78, 0.96]$, the necessity–sufficiency interaction is



(a) Oracle arms of known faithfulness.

(b) Citation fixed & the verdict mechanism is swept.

Figure 2: Positive-control validation. (a) IoU and paste-back necessity for oracle arms with known dependence structure. The non-causal oracle reports the same region as the faithful oracle but bases its verdict on a global cue. (b) In `mixed_cue`(β), the cited region is fixed while the verdict shifts from global to local evidence, leaving IoU constant while necessity changes.

Table 1: Positive-control validation on MagicBrush (400 audited rows per arm). Necessity measures the verdict response to restoring the cited region, while IoU measures localization. Intervals are row-level 95% bootstrap intervals.

oracle arm	necessity at $R(x)$	IoU
faithful ($\alpha = 1$)	0.991 $[0.991, 0.992]$	0.75 $[0.73, 0.76]$
global cue (correct, non-causal)	**0.00** $[0.00, 0.00]$	**0.75** $[0.73, 0.76]$
decoy	0.0002 $[0.0001, 0.0004]$	0.002 $[0.001, 0.003]$
half-evidence	0.992 $[0.991, 0.992]$	0.42 $[0.41, 0.42]$

nonzero, and the effect varies with coverage and weakens under feathering (Apx. E.6). We therefore interpret it only as verdict–output-region alignment under the specified construction.

All FakeShield effects use an elicited graded yes/no readout rather than the released prose. After full restoration, the prose changes to "authentic" on only 0 of 50 pairs, and forced-choice behavior varies by verbalizer. Thus, the reported effects characterize the specified readout, not the released paragraph itself.

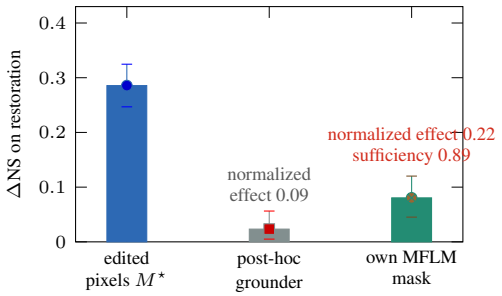
3.3.2 FFAA

FFAA provides a complementary case where the narration directly names coarse face regions. We use the released checkpoint and user question, with a teacher-forced `Analysis result` readout that agrees with the written verdict on 100.0% of images overall and 94.0% on the single-part corpus. The corpus replaces one face part—eyes, nose, or mouth—and is bit-identical outside the manipulated region.

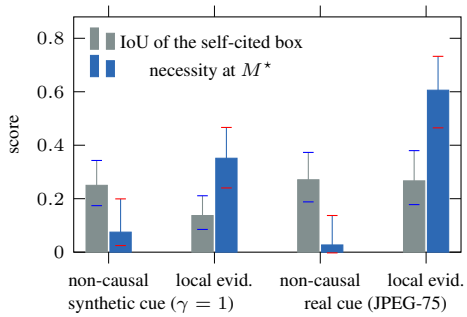
FFAA is a strong in-domain detector (AUC 0.997) and responds strongly to restoring the true manipulation ($\Delta\mathrm{NS} = +0.71$). Because no reliable box is available, we construct the intervention region directly from the face parts named in the narration. The model names the swapped part on 96.0% of images versus 54.0% under shuffled narrations. Restoring its own named parts changes the verdict score by +0.65, compared with +0.35 for shuffled narration, yielding an overall own-minus-shuffled effect of 0.30 $[0.23, 0.37]$ (Apx. E′).

This supports a narrow claim: FFAA's coarse part names identify regions more relevant to the specified verdict readout than unrelated narrations. It does not validate the full generated explanation or establish why the prose was produced. Moreover, its normalized effects should not be compared directly with FakeShield because the corpora, readouts, and region instruments differ.

Takeaway. FakeShield provides evidence about a system output mask, whereas FFAA permits a more direct narration-to-region intervention. In both cases, paste-back tests verdict–region alignment under the specified readout and region definition, not explanation faithfulness in general.



(a) FakeShield, certified corpus ($n = 50$).



(b) Trained arms: localization–intervention dissociation.

Figure 3: Released-system and learned-model audits. (a) FakeShield elicited yes/no changes at the ground-truth, post-hoc grounded, and native MFLM regions. (b) Learned arms on the original edit-ID holdouts; source exposure is not excluded.

Table 2: **Released-system audit summary.** Each effect is reported under its own readout and region instrument; values across systems are not directly comparable.

region restored	quantity	effect [95% interval]	n
FakeShield, certified splice; elicited yes/no readout			
edit $M^\star$	ΔNS	0.29 [0.25, 0.32]	50
post-hoc grounded region	ΔNS	+0.02 [+0.01, +0.03]	150
post-hoc grounded region	normalized necessity	0.09 [0.04, 0.16]	49
own MFLM mask	normalized necessity	**0.22** [0.12, 0.33]	49
own MFLM mask	normalized sufficiency	0.89 [0.78, 0.96]	49
FFAA, single-part faces; teacher-forced answer-field readout			
named parts minus shuffled narration	necessity	0.30 [0.23, 0.37]	150
named part masks	normalized necessity	0.99 [0.96, 1.00]	133

3.4 Learned-Model Stress Test

We next test whether the localization–intervention dissociation can emerge after training. MagicBrush provides a global re-encoding cue but weak local restoration, while our certified splice corpus provides exact local restoration but no global cue. We therefore add a controlled canvas-wide fingerprint and train three Qwen2-VL-2B LoRA arms with identical citation supervision, 120 steps, and edit-ID splits: `localonly`, `globalcue`, and `detect_cite`, which receives both cues. Because the arms solve different detection tasks, their difficulty and score calibration need not match. Table 3 summarizes the estimands, cohorts, and uncertainty units for the original and source-filtered analyses.

Synthetic global cue. At $\gamma = 1$, the global-cue arm localizes better than the local-only arm (IoU 0.25 $[0.17, 0.34]$ versus 0.137 $[0.085, 0.211]$). We therefore compare both at the same ground-truth mask $M^\star$, chosen by neither model. Despite near-perfect detection ranking, the global-minus-local necessity gap is -0.28 $[-0.40, -0.15]$. Thus, better localization can coexist with weaker restoration response in a trained VLM. Additional cue strengths, replications, and score-scale analyses appear in Apx. J.

JPEG-75 re-encoding cue. We repeat the design using whole-image JPEG-75 re-encoding (Wang et al., 2020; Frank et al., 2020). In seed 0 ($n = 40$), the global-minus-local gap is $\mathbf{-0.58}$ $[-0.71, -0.44]$ at the common ground-truth mask and -0.274 $[-0.400, -0.155]$ at the self-cited region. The IoU contrast, +0.0041 $[-0.0710, +0.0695]$ ($n = 40$), includes zero and establishes neither a difference nor equivalence. Detection AUCs are 0.997 and 0.87, so the arms are not response-matched. Across three edit-ID replications, both necessity gaps retain direction while the IoU contrast changes sign. The `detect_cite` arm lies between the designed arms in seed 0 but not consistently across splits (Apx. J.6).

Source-exposure sensitivity. Edit-ID splits do not ensure source independence because evaluation hosts can appear in training through other edits. Excluding every evaluation host or donor

Table 3: Current evidence and its estimands. Original and filtered rows reuse fitted models; changing the uncertainty unit does not create independent replications.

readout / region	estimand and cohort	uncertainty and supported scope
Learned score / gold mask	Original global-minus-local necessity; 40 edit-ID holdouts per seed	Paired row bootstrap; source-exposed construction. Stronger gold-region sensitivity below.
Learned score / own box	Original self-cited necessity; same 40 rows	Same row bootstrap; does not establish source-disjoint generalization.
Learned score / gold and own	D_R contrasts; original, host-unseen, strict 18/10/15 subsets	Host-source and joint host/donor component bootstrap. Gold intervals exclude zero; strict seed-2 own interval reaches zero.
Three fitted JPEG replications	Equal-weight per-seed means, original exposed cohorts	Three units, descriptive; older t_2 summaries are secondary, exact sign-test minimum $p = 0.25$.
FakeShield yes/no / MFLM mask	Necessity and sufficiency ratios, 49 eligible; raw interaction, 50	Ratios: original row bootstrap. Interaction: 47 host clusters; context dependence of elicited scores, not native prose.
FakeShield yes/no / MFLM mask	Feather-minus-hard: 48 complete cases; 46 shared-floor ratios	Host-cluster intervals; two failed masks undefined. Full-requested change bounds include zero. Shape and placement change in coverage controls.
FFAA teacher-forced field / named parts	Own-minus-shuffled necessity; 150 with empty-region identity	Generic LLaVA system wrapper; paired row bootstrap, strict lexicon and bounds. Coarse naming, not native-pipeline validation.

observed during training leaves 18/10/15 pairs across the three replications. We additionally use the calibration-invariant ranking diagnostic

$$D_R = \mathrm{AUC}(\text{target}, \text{source}) - \mathrm{AUC}(\text{restored}, \text{same source}),$$

which is unchanged by a common strictly increasing transformation of an arm's scores. On these reduced cohorts, the fixed-$M^\star$ global-minus-local gaps are -0.56, -0.50, -0.46, with all host-source-clustered intervals below zero. The self-cited gaps are -0.18, -0.20, -0.08; seed 2 reaches zero and therefore remains unresolved. A host/donor connected-component bootstrap preserves this distinction, although only 11/9/10 components remain (Apx. P.3). These are sensitivity analyses of the existing fitted models, not independent retraining experiments or evidence of population-level generalization.

Interpretation. This is a stress test, not mechanistic identification: paste-back removes both local edit content and any fingerprint within the restored region, and the global-cue model is evaluated on out-of-distribution restored inputs. Three replications also do not establish unseen-source generalization. Still, the synthetic and JPEG-75 results show that localization–intervention dissociation can arise in trained VLMs. The strongest evidence is at the common ground-truth region; source-filtered self-citation is supported in two replications and unresolved in one. Figure 3b shows the original edit-ID holdouts, with source/donor filtering reported as sensitivity analysis.

4 Discussion

What paste-back measures. Paste-back tests whether a specified verdict responds when the content inside a reported region is replaced by its authentic counterpart. It therefore measures *evidential necessity* under a particular intervention, not the model's complete internal computation. A region may participate in the computation without containing manipulation evidence, and restoration effects can interact with surrounding context. Accordingly, necessity should be interpreted together with localization rather than as a standalone measure of explanation faithfulness.

Interpreting system outputs. The released-system audits also show that the meaning of an intervention depends on how the region and verdict are obtained. FakeShield's native mask supports a claim about verdict–output-region alignment, but the mask is not established to depend on its narration. FFAA provides a more direct narration-to-region intervention because its text names the tested face parts, although this supports only coarse part-level alignment rather than the full generated ex-

planation. The two systems therefore illustrate different measurement settings and should not be compared through a single faithfulness score.

Scope and limitations. Paste-back requires a retained authentic source and is therefore unavailable for many fully generated or source-unknown images. Its conclusions are also specific to the chosen readout, region instrument, corpus, and intervention construction. Our released-system experiments are case studies rather than a survey of forensic VLMs, and the learned-model experiments are controlled stress tests rather than mechanistic identification. In particular, the strongest learned-model evidence comes from a common ground-truth region, while source-filtered self-citation remains unresolved in one replication. Broader claims will require more released systems and train/test splits that enforce source independence before training.

5 Related Work

Explainable forensic VLMs. Recent forensic VLMs augment manipulation detection with natural-language explanations, localization, or both (Xu et al., 2025b; Huang et al., 2024; Zhang et al., 2024a; Huang et al., 2025a; Wen et al., 2025; Kang et al., 2025; Sun et al., 2025; Zhou et al., 2025; Huang et al., 2025b; Jiang et al., 2026; Ji et al., 2026; Tan et al., 2026). Corresponding benchmarks evaluate explanation quality using reference-text similarity, evidence matching, localization overlap, or LLM-based judgments (Lin et al., 2025; Narang et al., 2026; Yang et al., 2025; Kuckreja et al., 2026). These evaluations measure whether an explanation is plausible or localized correctly, but they do not establish whether the model's verdict responds to the reported evidence. Our work addresses this complementary question for source-retained manipulations.

Faithfulness of model explanations. A long line of work distinguishes explanation plausibility from faithfulness: a model may produce an explanation that appears reasonable without using that evidence to reach its prediction (Ross et al., 2017; Jacovi & Goldberg, 2020; Turpin et al., 2023). Rationale-based evaluation in NLP has therefore used comprehensiveness and sufficiency interventions, while vision methods evaluate saliency through deletion, insertion, occlusion, and related perturbations (DeYoung et al., 2020; Petsiuk et al., 2018; Adebayo et al., 2018). These methods motivate our use of intervention response alongside localization. The forensic setting, however, makes the intervention itself unusually important: masking, deletion, or synthesis can introduce traces that a forensic detector may treat as evidence.

Counterfactual and intervention-based evaluation. Counterfactual explanation methods replace or remove visual evidence and measure the resulting prediction change (Chang et al., 2019; Goyal et al., 2019; Kaushik et al., 2020). The closest prior work is EDCT (Ding et al., 2025), which treats a VLM explanation as a falsifiable hypothesis, identifies concepts from the explanation, removes them using generative inpainting, and measures changes in the answer and explanation. Our necessity quantity is closely related to EDCT's answer-change component; we therefore do not claim answer change itself as novel. The distinction is the intervention: for source-retained forensic edits, the authentic pre-edit image supplies observed replacement pixels, so paste-back requires no synthesized counterfactual and is exactly the identity on authentic inputs. We further validate this intervention with correct-but-non-causal positive controls that keep localization fixed while changing verdict dependence. Thus, our contribution is the held-source intervention and its validation protocol for forensic auditing, rather than a new scalar faithfulness metric.

6 Conclusion

We introduced *paste-back auditing*, a held-source intervention for testing whether a forensic VLM's verdict responds to its reported region. By restoring authentic source pixels, paste-back avoids introducing synthesized counterfactual content and is exactly the identity on authentic inputs. Our positive controls show that localization and intervention response capture distinct properties, and experiments on released and trained VLMs show that plausible localization need not imply strong verdict dependence on the localized evidence. These results support evaluating forensic explanations with localization and intervention response together, while stating the readout and region instrument explicitly.

REPRODUCIBILITY STATEMENT

Every number in the tables and figures, and every macro-expanded number in the prose, is produced by the released code from stored rows; the commands that reproduce each stage are listed in Appendix N. **Sandbox results run fully offline** (no network, no weights, no CUDA): the test suite, the end-to-end pipeline (about three minutes) and this document's build need nothing else. **The primary audit** loads `zhipeixu/fakeshield-v1-22b` at revision `db861d30` into a native LLaVA container by a bijective 758-tensor rename, gates the load on bitwise equality with the safetensors bytes, and records the resulting load report in every results row; MFLM is loaded the same way (968 tensors, 0 missing, 27/27 fingerprints bitwise equal) and additionally cross-checked against Meta's reference SAM. It costs 1482 local forward passes and 57 minutes on one Apple-silicon machine with no network and no cache, so it is reproducible but not free; the $n = 150$ extension costs a further 4419 forward passes and 189 minutes, and a stored-row certificate proves that it contains the baseline $n = 50$ run and reproduces it bit-exactly before any of its numbers are read. **End-to-end implementation equivalence.** The refactored loader passes the identical bitwise gate ($10/10$ checkpoint fingerprints) — but a fingerprint gate establishes only that the selected weights are the same bytes. It does *not* establish that preprocessing, prompt construction, tensor routing or the native generated text are unchanged, and those are what a verdict is read through. So the audit was re-run at $n = 50$ through the refactored code (`scripts/queues/round10_loader_audit.sh`; 1482 forward passes, $\approx$ 57 minutes, no network). It reproduces the baseline rows **exactly**: 537 of 537 stored numeric cells across 11 columns are bit-equal (largest absolute difference 0), and the free-text outputs are identical on 50 of 50 explanations and 50 verdict strings — which is the part no fingerprint gate can reach. The re-audit writes only to `*_postrefactor` files, so the baseline rows remain available for comparison, and its stage post-conditions assert that the comparison *happened*, never that it agreed: a disagreement was an outcome this run was built to be able to report. **Served-model results** need the MagicBrush download and a local Ollama server; one configuration file pins models, prompts, seed, grid, control rules and a hard 6000-request budget enforced inside the HTTP transport, all responses are cached on request bytes, and re-running the complete 150-row design cost 115 of 4350 requests and reproduced every reported number. **The positive control and the stratification** need the download and nothing else ($\sim$30 CPU minutes). **White-box results** need `torch`, `transformers`, `peft` and one GPU/MPS device. Aggregation is separate from inference: statistics are reconstructed from stored rows, with explicit handling of paired finite values and exact-zero outcomes. Adapters for FaceForensics++, CASIA v2, IMD2020, AutoSplice, DocTamper and the Stable-Diffusion-inpainting and MAT operators are implemented, documented and **were not run** (the learned big-lama operator **was** run; and the registry key `inpaint_diffusion` names a classical Laplace fill, not a denoising-diffusion model — Appendix N).

ETHICS STATEMENT

This work is an audit. It measures two released forensic checkpoints (FakeShield with its MFLM localiser, and FFAA), three locally served open-weight chat VLMs, and LoRA-trained research arms on a Qwen2-VL-2B we host, on MagicBrush, on splice/copy-move corpora we construct from public images, and on two face corpora we construct from CelebAMask-HQ; no other data were collected. The face corpora composite one CelebA-HQ portrait's eyes, brows, nose and mouth into another's under that release's non-commercial research terms, which is the most ethically salient construction here: they are region-replacement forgeries of real people's photographs, built only because auditing a face-forensics model requires a manipulation whose authentic source is retained. They are constructed by `scripts/build_face_*_corpus.py` from images we do not redistribute, they depict no one in a false context beyond the swap itself, they use a documented hard-paste construction, and they are intended only for detector evaluation. This intent does not eliminate potential misuse. Paste-back writes an image's own authentic source pixels back into it and invokes no generator, so the audit creates no new *generated* content, and every model call is against a checkpoint we hold or a server we run. We consider publication net-positive: the failure mode we measure — fluent narration or localisation that is not tied to the intervened pixels — is already visible to anyone who compares a verdict with a restored image, and the audit gives a defender a way to quantify the exposure rather than a new way to exploit it. We release no adversarial examples and ran no attack against any external service. The two explanation-aware threat models sketched in Appendix L.1 are a synthetic-sandbox study (four small models, sixteen examples) that supports no claim about any

deployed system, and no white-box adaptive attack was evaluated on a real model (Appendix K); the closest released attack on VLM-backbone detectors is prior work (Li et al., 2026).

AI USE STATEMENT

Generative-AI assistants were used for code debugging. Reported numbers are not generated by those assistants: they are produced from stored experiment rows by the cited repository scripts and checked with automated tests, artifact-freshness gates, and rendered-PDF verification. Model-based components used as experimental instruments (including grounders and rubric judges) are identified in the corresponding methods and results. The authors take responsibility for all content.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

Belhassen Bayar and Matthew C. Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, 2016.

Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white- and black-box attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Isha Chaturvedi, Anjana Nair, Yushen Li, Adhitya Rajendra Kumar, Kevin Zhu, Sunishchal Dev, Ashwinee Panda, and Vasu Sharma. Peek-a-boo reasoning: Contrastive region masking in MLLMs. *arXiv preprint arXiv:2512.08976*, 2025.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Sihao Ding, Santosh Vasa, and Aditi Ramadwar. Explanation-driven counterfactual testing for faithfulness in vision-language model explanations. In *Proceedings of the Proceedings of the Advances in Neural Information Processing Systems Workshop on Regulatable ML (NeurIPS)*, 2025.

Shichao Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. Explaining deepfake detection by analysing image matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

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 International Conference on Machine Learning (ICML)*, 2020.

Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3), 2012.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Zhengchao Huang, Bin Xia, Zicheng Lin, Zhun Mou, Wenming Yang, and Jiaya Jia. FFAA: Multimodal large language model based explainable open-world face forgery analysis assistant. *arXiv preprint arXiv:2408.10072*, 2024.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a.

Zhenglin Huang, Xiangtai Li, Xi Yang, Bei Peng, Xiaowei Huang, Baoyuan Wu, Dacheng Tao, Ming-Hsuan Yang, and Guangliang Cheng. So-Fake: Benchmarking and explaining social media image forgery detection. *arXiv preprint arXiv:2505.18660*, 2025b.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

Yikun Ji, Yan Hong, Qi Fan, Jun Lan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, and Jianfu Zhang. FakeXplain: AI-generated image detection via human-aligned grounded reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Changjiang Jiang, Wenhui Dong, Zhonghao Zhang, Fengchang Yu, Wei Peng, Xinbin Yuan, Yifei Bi, Ming Zhao, Zian Zhou, Chenyang Si, and Caifeng Shan. Ivy-Fake: A unified explainable framework and benchmark for image and video AIGC detection. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, 2026.

Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, and Conghui He. LEGION: Learning to ground and explain for synthetic image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Kartik Kuckreja, Parul Gupta, Muhammad Haris Khan, and Abhinav Dhall. Pixels don't lie (but your detector might): Bootstrapping MLLM-as-a-judge for trustworthy deepfake detection and reasoning supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Haipeng Li, Rongxuan Peng, Anwei Luo, Shunquan Tan, Changsheng Chen, and Anastasia Antsiferova. Universal anti-forensics attack against image forgery detection via multi-modal guidance. *arXiv preprint arXiv:2602.06530*, 2026.

Yang Li, Songlin Yang, Wei Wang, Ziwen He, Bo Peng, and Jing Dong. Counterfactual explanations for face forgery detection via adversarial removal of artifacts. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2024.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Junyan Ye, Ke-Yue Zhang, Yue Zhou, Peng Jin, Bin Li, Taiping Yao, and Shouhong Ding. Seeing before reasoning: A unified framework for generalizable and explainable fake image detection. *arXiv preprint arXiv:2509.25502*, 2025.

Weijiang Lv, Yaoxuan Feng, Xiaobo Xia, Jiayu Wang, Yan Jing, Wenchao Chen, and Bo Chen. SPD-Faith Bench: Diagnosing and improving faithfulness in chain-of-thought for multimodal large language models. *arXiv preprint arXiv:2602.07833*, 2026.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Johannes Moll, Markus Graf, Tristan Lemke, Nicolas Lenhart, Daniel Truhn, Jean-Benoit Delbrouck, Jiazhen Pan, Daniel Rueckert, Lisa C. Adams, and Keno K. Bressem. Evaluating reasoning faithfulness in medical vision-language models using multimodal perturbations. In *Proceedings of the Conference on Machine Learning for Health (ML4H)*, 2025.

Abhijeet Narang, Kartik Kuckreja, Shreya Ghosh, Muhammad Haris Khan, Jianfei Cai, and Abhinav Dhall. XPlainVerse: A million-scale benchmark for explainable deepfake detection. *arXiv preprint arXiv:2607.03562*, 2026.

Van-Loc Nguyen, AprilPyone MaungMaung, Minh-Triet Tran, and Isao Echizen. EditSleuth: A dataset of grounded reasoning chains for image-edit forensics. *arXiv preprint arXiv:2605.08695*, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Letitia Parcalabescu and Anette Frank. Do vision & language decoders use images and text equally? how self-consistent are their explanations? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

Laura Rieger, Chandan Singh, W. James Murdoch, and Bin Yu. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

Zhihao Sun, Haoran Jiang, Haoran Chen, Yixin Cao, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. ForgerySleuth: Empowering multimodal large language models for image manipulation detection. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Maximilian Süss. Faithfulness of natural language explanations for vision language models: An automated test framework. Diploma thesis, Technische Universität Wien, 2026.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022.

Hao Tan, Jun Lan, Zichang Tan, Ajian Liu, Chuanbiao Song, Senyuan Shi, Huijia Zhu, Weiqiang Wang, Jun Wan, and Zhen Lei. Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of Graphics Tools*, 9(1), 2004.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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 (CVPR)*, 2020.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yao Xiao, Weiyan Chen, Jiahao Chen, Zijie Cao, Weijian Deng, Binbin Yang, Ziyi Dong, Xiangyang Ji, Wei Ke, Pengxu Wei, and Liang Lin. Unveiling perceptual artifacts: A fine-grained benchmark for interpretable AI-generated image detection. *arXiv preprint arXiv:2601.19430*, 2026.

Tianrun Xu, Haoda Jing, Ye Li, Yuquan Wei, Jun Feng, Guanyu Chen, Haichuan Gao, Tianren Zhang, and Feng Chen. DeFacto: Counterfactual thinking with images for enforcing evidence-grounded and faithful reasoning. *arXiv preprint arXiv:2509.20912*, 2025a.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025b.

Mengjiao Yang and Been Kim. Benchmarking attribution methods with relative feature importance. *arXiv preprint arXiv:1907.09701*, 2019.

Michael Yang, Shijian Deng, William T. Doan, Kai Wang, Tianyu Yang, Harsh Singh, and Yapeng Tian. Explainable AI-generated image detection RewardBench. *arXiv preprint arXiv:2511.12363*, 2025.

Fanrui Zhang, Jiawei Liu, Jiaying Zhu, Esther Sun, Dong Li, Qiang Zhang, and Zheng-Jun Zha. ForgeryGPT: A multimodal LLM for interpretable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*, 2024a.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. MagicBrush: A manually annotated dataset for instruction-guided image editing. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024b.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. AIGI-Holmes: Towards explainable and generalizable AI-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

How to read this appendix

This appendix provides detailed methods, tables, controls and negative results. Supplementary sandbox and served-model analyses are summarised in §L. The machine-generated manifest below is the authoritative map from each load-bearing claim to its model, corpus, region instrument, stored rows and minimal regeneration command; a missing input is printed as *pending*, never treated as support. The source-exposure, ranking and paired-robustness diagnostics in §P qualify the evaluation cohorts and are linked from `results/claim_manifest.json`. Table 3 is the main-text estimand map; `make final-results` emits the source/dependence, inclusion, readout and lexicon export. Read §P, including the component sensitivity, before interpreting the older learned tables.

Contents of the appendix.

How to read this appendix **15**

A Related work **19**

B Study design and evidence scope **22**

C Method in full **23**

C.1 Setup . . . 23
C.2 Intervention and localisation quantities . . . 24
C.3 Two refinements that answer the obvious objections . . . 26
C.4 Grounding . . . 26
C.5 Mechanistic decoupling . . . 27
C.6 FaithAlign . . . 27
C.7 Threat models . . . 28

D The validation protocol in full: a positive control with a discriminating arm **28**

E The primary audit in full: a released explainable forensic VLM **33**

E.1 Detection first . . . 34
E.2 Local sensitivity and an unresolved citation effect on MagicBrush . . . 34
E.3 Audit on a certified local-manipulation corpus . . . 35
E.4 Is the citation-null falsifiable? A positive control for the grounder . . . 38
E.5 Swapping the instrument: a served VLM as the grounder . . . 39
E.6 Dissolving the grounder confound: the system's own localiser . . . 40
E.7 An adjudicated citation subset, and who adjudicated it . . . 45
E.8 The null, bounded by power rather than asserted . . . 46
E.9 Conditioning on detection . . . 47
E.10 Against the general-purpose models it is supposed to beat . . . 47
E.11 Text-metric associations on the released forensic system . . . 48
E.12 Composite-score redundancy . . . 48
E.13 Outside-mask residuals on MagicBrush . . . 48
E.14 Native prose versus forced-choice readout . . . 50
E.15 Threats specific to this section . . . 51

E′ The second released system: FFAA on a face corpus that holds the counterfactual **51**

F The secondary contrast: three general-purpose chat VLMs **56**

G Sensitivity to the counterfactual operator **58**

H Inside a real VLM: activation patching and FaithAlign **61**

H.1 FaithAlign: what a counterfactual training term buys 62

H.2 The hinge objective and the controlled contrast 64

H.3 The global-cue contrast, reproduced on a trained network 64

H.4 Activation patching: does the narration move with the verdict? 68

I Separating the pathology from the corpus **70**

J The hybrid corpus in full **75**

J.1 The resolution–global-cue trade-off 75

J.2 Manufacturing the cue at a known strength 75

J.3 Result: the fingerprint sweep 76

J.4 The jointly supervised detector also uses the global cue 79

J.5 Replication across seeds 80

J.6 The same design with a real cue: the JPEG-75 replication 80

J.7 What the hybrid corpus does not settle 82

K Limitations and scope **82**

K.1 Sensitivity to corpus, estimator and instrument 89

L Supplementary sandbox and served-model analyses **90**

L.1 The synthetic sandbox and its four experiments (E1–E4) 90

L.2 The served-VLM audit on MagicBrush 91

M The real corpus: MagicBrush preparation and alignment gate **91**

N Implementation notes **92**

O Limitations of a composite faithfulness score **93**

P Source exposure, score scale, and paired robustness **94**

P.1 Edit-ID holdout and source-image exposure 94

P.2 A monotone-invariant intervention diagnostic 95

P.3 Joint host/donor construction dependence 95

P.4 Paired feathering and missing outputs 96

P.5 Context dependence from bidirectional interventions 97

Analysis-choice ledger. There is no untouched validation subset available for decisions already made; we therefore do not retroactively call them preregistered. The ledger states which choices were initial, revised after inspection, or added as sensitivity analyses.

Table 4: **CLAIM-TO-EVIDENCE MANIFEST.** Row counts and SHA-256 fingerprints are also stored in `results/claim_manifest.json`; regenerate both artifacts with `make claim-manifest`.

Claim	model / corpus	region instrument	status	row n	minimal command
necessity separates faithful and global-cue oracles at identical IoU	oracle families / MagicBrush paired source	oracle ground-truth box	complete	6000	`make positive-control`
FakeShield post-hoc citation effect on hard-paste corpus	FakeShield DTE-FDM / splice/copy-move hard paste	post-hoc text grounder	complete	150	`make fakeshield-splice150-analyze`
FakeShield output-region intervention on hard-paste corpus	FakeShield DTE-FDM + MFLM / splice/copy-move hard paste	native MFLM output mask	complete	50	`make mflm-audit`
hard-paste normalized intervention effect with floor sensitivity	FakeShield DTE-FDM + MFLM / splice/copy-move hard paste	native MFLM output mask	complete	50	`make mflm-citation`
FakeShield output-region robustness to feathered construction	FakeShield DTE-FDM + MFLM / splice/copy-move feathered 3 px	native MFLM output mask	complete	48	`make mflm-feather`
FFAA uses parts named by its narration beyond shuffled narration	FFAA Mistral-7B / single-part CelebAMask-HQ replacement	union of part masks named in narration	complete	150	`make ffaa-parts`
trained score-localisation dissociation on edit-ID holdouts; source independence unestablished	Qwen2-VL-2B LoRA arms / hybrid analytic fingerprint gamma=1	M* and narration-embedded box	complete	160+80+80	`make hybrid-analyze`
trained dissociation under JPEG-75 on source-exposed edit-ID holdouts; strict seed-2 self-citation unresolved	Qwen2-VL-2B LoRA arms / hybrid JPEG-75 re-encode	M* and narration-embedded box	complete	160+120+120	`make final-results`
forced-choice readout agreement with native verdict before/after restoration	FakeShield DTE-FDM / splice/copy-move hard paste	target and full restored M*	complete	50	`make fakeshield-readout`
necessity and sufficiency on the same native MFLM masks	FakeShield DTE-FDM + MFLM / splice/copy-move hard paste	deterministically reconstructed native MFLM output mask	complete	50	`make mflm-sufficiency`

Table 5: **CONSOLIDATED ANALYSIS LEDGER.**

choice	first run	analysis status	observation or procedure	interpretation
MFLM absolute box containment $\geq .90$	2026-08-16	criterion not met	Containment was below the chosen cutoff.	The randomised-decoder contrast was selected after inspection and is exploratory.
MFLM 8-member multiplicity family	2026-08-16	frozen before native-versus-grounder contrast	All eight members named in `manifest.family_for_bh`, including nulls, enter one BH family, and all eight are printed in Table 9.	Means, rank tests, and corrected q values are all reported; no relabelling as confirmatory validation.
White-box raw probability contrast	2026-08-09	initial analysis	Mean was carried by two rows near probability saturation.	Raw result retained; log-odds was added post hoc and is labelled sensitivity, not promoted to a preregistered endpoint.
Hybrid seed expansion	2026-09-08	post-inspection replication	Extra seeds use new permutations and fingerprints.	Seed/split is the replication unit; duplicate fingerprints are rejected by code.
JPEG IoU row handling	2026-09-08	complete-pair analysis	Exact-zero IoUs are valid observations, not missing values.	All 40 seed-0 pairs are retained; the gap is $+0.0041$, and non-significance is not called equivalence.
Denominator floor grid	2026-09-08	exploratory sensitivity	Primary floor is retained and a 0.01–0.10 grid is added.	Every floor reports kept/dropped n; the ratio is a normalized intervention effect, not evidence allocation.
Feathered-construction repeat of the native-mask audit	2026-09-08	exploratory robustness	Requested the same 50 ids on the 3-px-feathered build.	Reported as its own cohort with its own floor grid; it is not pooled with the hard-paste rows.
Bidirectional same-mask interventions	2026-09-08	exploratory analysis	Sufficiency added on the identical reconstructed native masks.	Necessity and sufficiency are printed together; neither is read as an evidence share.
Native-prose versus forced-choice readout	2026-09-08	exploratory validation	Three verbalizers scored before and after full restoration.	Agreement and its Wilson intervals are printed per verbalizer; the audited quantity stays the elicited readout.
MFLM ratio against mask coverage	2026-09-11	post-inspection analysis	Per-row normalized necessity read against each native mask's share of $M^{\star}$; Spearman primary, median split and three fixed cuts printed.	The reading is geometric; all cuts are reported without selecting a preferred cut.
JPEG-75 pooled across seeds	2026-09-11	exploratory analysis	Replication-level pool of per-seed paired means, t on 2 df, variance floored at the within-replication one.	The per-seed table stays primary; with 3 replications no replication-level randomization p can fall below 0.05.

Reading Table 5. "Frozen" applies only from the point named in the row; a revised criterion is not described as originally predeclared. "First run" is the UTC date stamped inside the result file each row is about, so the order of decision and inspection is checkable.

	Appendix	Why you would read it
Load-bearing: every claim in the main text rests on one of these.		
D	§D, the positive control	All 10 oracle arms and both sweeps, testing whether the audit separates known dependence mechanisms.
E	§E, the primary audit	The released system: the load gate, the certified-corpus replication (§E.3, including the $n = 150$ extension and its bit-exactness certificate), the grounder controls (§E.4) and the system's own localiser (§E.6).
E′	§E′, the second released system	FFAA through the identical instrument on two face corpora that hold the counterfactual: the readout its released format forces, the part-name citation test with a shuffled-narration control, the sufficiency direction for both systems, and the other nine released paraphrases.
J	§J, the hybrid corpus	Every arm at every γ, every paired contrast with its BH q, the seed replication, and the JPEG-75 real-cue run with the arms' self-cited necessities (§J.6).
P	§P	Required source-exposure, host/donor-component, ranking, missing-output and interaction sensitivities of the existing fitted models.
K	§K, limitations in full	Construct validity, model and corpus coverage, statistical uncertainty, and sensitivity to measurement choices.
Supporting: methods and secondary evidence.		
A–C	§A–§C	Related work, the full contribution list with the boundary between our kinds of evidence, and the method in full.
F–I	§F–§I	The chat-model contrast; why the choice of counterfactual operator determines your conclusion; the white-box activation-patching and FaithAlign runs on a model we host; and the corpus/pathology separation.
L	§L	The synthetic sandbox's four experiments and the served-VLM audit, with their scope and limitations. These are supplementary evidence, not the basis for claims about released forensic systems.
M–O	§M–§O	Corpus preparation and its alignment gate, implementation notes with every reproduction command, and the limitations of a composite score.

Key interpretive limitations are the outside-mask residual (§E.13), outlier sensitivity of white-box means (§H.3.1), and redundancy of the composite score with IoU (§O). Section K.1 summarises how corpus, estimator and instrument choices constrain the conclusions.

A Related work

Explainable forensic VLMs — and which of them we actually run. FakeShield (Xu et al., 2025b), FFAA (Huang et al., 2024), ForgeryGPT (Zhang et al., 2024a) and SIDA (Huang et al., 2025a) produce narrated verdicts, and DD-VQA (Zhang et al., 2024b) supplies human explanations for face forgery. The population has grown through 2025–26: FakeVLM (Wen et al., 2025), LEGION (Kang et al., 2025), ForgerySleuth (Sun et al., 2025), AIGI-Holmes (Zhou et al., 2025), So-Fake (Huang et al., 2025b), Ivy-Fake (Jiang et al., 2026), FakeXplainer (Ji et al., 2026) and Veritas (Tan et al., 2026) all narrate, ground or segment the artifact they decide on, and the benchmarks built to grade such explanations — ExplainFake-Bench (Lin et al., 2025), XPlainVerse (Narang et al., 2026), XAIGID-RewardBench (Yang et al., 2025) and DeepfakeJudge (Kuckreja et al., 2026) — score them by overlap with a reference, by entity/evidence matching, or by an LLM judge. All of these are functions of the generated text and the annotated region; none asks whether the narrated artifact causally drives the verdict, and none trains for it. Most of the 2025–26 systems operate on whole-image generation, where no authentic source exists and paste-back is undefined; the comple-

mentary finding there is X-AIGD's (Xiao et al., 2026), that existing generated-image detectors rely negligibly on the perceptual artifacts they are meant to explain. EditSleuth (Nguyen et al., 2026) is the nearest public corpus to ours: edited/source/mask triplets with reasoning chains on the Qwen2-VL-2B backbone we host, on which paste-back applies unchanged. Our focus is measurement rather than a broadly validated training remedy.

Because that indictment is only worth as much as the population it is tested on, we state exactly which of these we obtained and ran. **FakeShield we ran.** Its weights are public, ungated and Apache-2.0 (`zhipeixu/fakeshield-v1-22b`), and Section E audits its detection-and-explanation module through the same instrument as every other model in this paper. Two implementation details matter for interpreting that audit. First, "v1-22B" is the size of the release, not of the narrating model: the DTE-FDM module is a LLaVA-1.5-13B-class network (13.351 B parameters), and the remaining capacity is MFLM, a separate GLaMM/SAM localiser. Second, DTE-FDM emits *no* grounding-token box, under any prompt we tried (0% of 50 audited rows and 0 of 25 in a dedicated prompt probe). Localisation in that system lives entirely in MFLM. Anyone citing FakeShield as a box-emitting explainable detector should say which module they mean. **FFAA we ran (Appendix E′); ForgeryGPT and SIDA we did not.** The weights of FFAA, FakeVLM, SIDA are public and ungated under the identifiers their own repositories link (`thu-huangzc/ffaa-mistral-7b`, `lingcco/fakeVLM`, `saberzl/SIDA-7B`, `saberzl/SIDA-13B`, `saberzl/SIDA-7B-description`, `saberzl/SIDA-13B-description`; 6 of 6 resolving identifiers neither gated nor private on 2026-09-05), so for those systems the gap is effort, not availability; ForgeryGPT's do not resolve under any identifier we found. Nothing measured on FakeShield transfers to them by assumption. This is the paper's main remaining gap in breadth and Section K names it as such.

Rationale faithfulness in NLP. ERASER (DeYoung et al., 2020) formalised comprehensiveness and sufficiency by deleting the rationale from the input; Turpin et al. (2023) showed chain-of-thought explanations can be systematically unfaithful; Jacovi & Goldberg (2020) argued for evaluating faithfulness separately from plausibility – the distinction this paper operationalises for vision. The porting is not mechanical. ERASER-style deletion in an image means content-free erasure, which introduces a new artifact a forensic model reacts to; our counterfactual restores held authentic pixels and adds no synthesis trace. Section L.1 reports comprehensiveness and sufficiency alongside necessity to quantify the difference.

Saliency faithfulness in vision. RISE (Petsiuk et al., 2018) and deletion/insertion curves, the pointing game, and the sanity checks of Adebayo et al. (2018) audit *heatmaps*. They share our concern but not our object: a heatmap has no semantics to be unfaithful about, whereas a sentence makes a specific claim about a specific artifact. They also use perturbations rather than held-source restorations, and have no notion of a suspicion-matched control. We implement RISE deletion and insertion AUC as baselines, plus a *restoration* variant that walks the same ordering with A_R instead of an erasure, so the two interventions can be compared on identical region orderings.

Counterfactual explanation testing. EDCT (Ding et al., 2025) is the closest prior work and the sharpest threat to our novelty. It is one year old, it already treats a VLM's natural-language explanation as a falsifiable hypothesis, and it already does the core move: parse the explanation into concepts, inpaint those concepts away, and score a Counterfactual Consistency Score in an audit framing. **We state the delta precisely, because a vaguer statement would flatter us.** CCS scores changes in the answer *and* in the explanation under LLM-assisted analysis; our necessity leg is the answer-change half, in a forensic setting, and we implement CCS's answer-change leg as a baseline (`rightpixels.cfscore.edct`). **Provenance, stated once and consistently.** We verified the EDCT paper directly (arXiv:2510.00047, NeurIPS 2025 Workshop on Regulatable ML): CCS has an answer-change leg and an explanation-change leg, and its counterfactual is built by generative inpainting of concepts parsed from the explanation. We found no public implementation, so ours is a reimplementation from the paper's own description of the method; where the paper does not fix a threshold we say which value we chose. What is genuinely ours is: (i) a *held* rather than synthesised counterfactual, whose false-flip rate on authentic content is zero by construction rather than by measurement — EDCT's is necessarily generative; (ii) the false-flip calibration itself, and the finding that the size of the inpainting confound is a property of the *detector* rather than of the inpainter (Section G), which is a result about EDCT-style protocols in general; (iii) a positive con-

trol with a discriminating correct-but-non-causal arm, which is what separates a causal metric from a localisation metric (Section D). The suspicion-matched control is not a distinct contribution: its advantage remains unresolved in three control evaluations (sandbox, n=12, n=50), so it is reported as a null and kept only for its residual use as an operator-trace estimator. We run CCS on our setting and report its spread across operators next to CF-Score's stability.

The positive control validates the protocol, not a unique score. On the positive control's decisive cell — the correct-but-non-causal global-cue arm — CCS's answer leg computed with *our* operator separates the arms as completely as CF-Score does (1.000 against 0.000, gap 1.000 $[1.000, 1.000]$). So the composite's advantage on that contrast is over *EDCT-with-a-synthesised-counterfactual*, not over EDCT's criterion as such; what CF-Score adds on top is resolution (CCS saturates on 60% of the citation sweep) and the incomplete-evidence cell that a binary criterion cannot see (1.000 against 1.000 where CF-Score reads 0.4097 against 0.7322). We therefore state the delta to EDCT as (i) the held counterfactual, (ii) the calibration, and (iii) the positive control as a validation *requirement* — and not as a claim that the composite makes a contrast CCS cannot.

Positive controls with known-by-construction ground truth. The protocol we ship — build systems whose faithfulness is known before the metric is applied, audit them all with one measurement procedure, and require the metric to recover the designed ordering — is *not* new as a move, and it needs naming. BAM (Yang & Kim, 2019) constructs a dataset and a family of models whose *relative* feature importance is known a priori, precisely so that attribution methods can be ranked against something other than human plausibility. ROAR (Hooker et al., 2019) is the canonical statement of the objection our operator section measures: removing the features an estimator calls important also introduces a distribution shift, so the measured degradation confounds the estimator's quality with the artefact of removal — ROAR's answer is to retrain after removal. The deltas we claim are narrow and we state them: BAM's ground truth is relative feature importance in an image classifier, whereas ours is the *causal structure of a narrated verdict* (an arm that is accurate, cites the right box, and is driven by something else is not expressible in BAM's design); and ROAR's repair, retraining on the ablated distribution, is unavailable for a served VLM one can only query — our repair is instead to hold the counterfactual rather than synthesise it, which writes no new synthesis trace and is the exact identity when the restored image is the authentic source. That is narrower than calling the intervention distribution-preserving: a *partially* restored manipulated image is not guaranteed to be a natural sample, and boundaries, restored–surround interaction and redundant cues remain (§C). Neither delta makes the protocol novel; what it makes it is *correct*, which is why we ship it as a requirement rather than as a contribution.

Held versus synthesised counterfactuals. FIDO (Chang et al., 2019) already argues that content-free erasure (blur, noise, grey) produces inputs off the data manifold and that a generative infill conditioned on the rest of the image is the right intervention; it is the direct ancestor of our Section G, and the operator it advocates is exactly the one we measure as sign-inverting on some detectors. Counterfactual Visual Explanations (Goyal et al., 2019) already swaps regions *from a donor image* to ask what would have to change for the class to change. Kaushik et al. (2020) already makes "the counterfactual is elicited, not synthesised" the central methodological point, in NLP, by having humans write minimally revised documents. In forensics itself, Dong et al. (2022) already use the fake/source/target triplet a face-swap corpus retains to explain what a detector has learned to match — the closest prior use of the held source, as a training-set analysis rather than as an intervention on a narrated verdict — and Li et al. (2024) synthesise counterfactuals for a face-forgery detector by adversarially removing artifacts in a StyleGAN latent, which is exactly the synthesised-counterfactual confound Section G measures. Our Proposition 2 is, honestly stated, that manipulation forensics is a setting where Kaushik et al.'s move is *free*: the corpus already ships the minimally revised document, so the infill FIDO must learn and Goyal et al. must borrow is simply read off disk, and the false-flip rate that both must measure is zero by construction. That is a domain observation with a two-line proof, not a new mechanism, and we do not present it as one.

Faithfulness of free-text explanations, including for VLMs. Atanasova et al. (2023) test natural-language explanations by constructing counterfactual inputs whose predictions flip and asking whether the explanation reflects the inserted reason — the same logical shape as our necessity leg, in NLP and with a synthesised edit. CC-SHAP (Parcalabescu & Frank, 2024) measures self-consistency between the input contributions that drive the answer and those that drive the explana-

tion, and its VLM variant (Parcalabescu & Frank, 2025) makes it the closest measurement instrument to ours in modality. Region-level interventions on VLM explanations have since appeared in several forms, every one of which synthesises or erases the counterfactual: Contrastive Region Masking (Chaturvedi et al., 2025) masks annotated regions and contrasts the reasoning trace with the unmasked one and with a random region masked at least five per cent away; DeFacto (Xu et al., 2025a) trains with counterfactual and random-masked images to enforce evidence-grounded answers; SPD-Faith (Lv et al., 2026) and the medical-imaging tests of Moll et al. (2025) perturb the input and read the chain of thought; and the vCT/vCCT framework (Süss, 2026) removes objects and correlates the verdict shift with whether the explanation mentions the removed object. The difference that matters is what plays the role of ground truth: CC-SHAP compares two attributions of the *same* model to each other and is therefore silent when both are wrong in the same way, whereas paste-back compares the model's verdict against an intervention on the world whose correct answer we know. The cost of that is that our instrument only exists where the authentic counterpart is held, which is a much narrower setting than CC-SHAP's.

Global generator cues. That an accurate generated-image detector may be reading a canvas-wide generator signature rather than any local evidence is the founding observation of the generated-image-detection literature, not a discovery of ours: Wang et al. (2020) show a detector trained on one generator transfers across architectures, and Frank et al. (2020) identify the frequency-domain artefacts that make it possible. We use that fact as an *instrument*: our hybrid corpus manufactures such a signature at a known strength γ (Section J) so that a correct-but-non-causal detector can be trained on demand and its necessity measured, which is what turns a known phenomenon into a controlled independent variable.

Right for the right reasons and attention supervision. Ross et al. (2017) and Rieger et al. (2020) constrain input gradients towards an annotated mask. FaithAlign's grounding loss is that idea, and our ablation treats it as the baseline to beat rather than as a contribution: Section L.1 shows it buys localisation and no faithfulness at all. The counterfactual-consistency and attribution-alignment losses are causal (they constrain the effect of an intervention, not the value of a gradient at a point) and are intended to align the language citation with the verdict pathway; our white-box experiments test, but do not generally establish, that alignment.

Causal mediation and activation patching. Causal tracing and path patching (Meng et al., 2022; Wang et al., 2023) localise computation by intervening on activations. We adapt the technique to the visual-token $\rightarrow$ {verdict head, citation head} pathway. We use patching rather than attention rollout deliberately: rollout is correlational, and Jain & Wallace (2019) and the sanity-check literature give ample reason not to read it as a causal account.

Attacks on forensic detectors. Adversarial evasion of deepfake and splice detectors is well studied; Carlini & Farid (2020) established the methodological standard — attack the whole system, attack the defended model, report budgets — and Li et al. (2026) give a universal anti-forensics attack against VLM-backbone forgery detectors, the systems this paper audits. Prior work attacks the *verdict*. The two explanation-aware threat models sketched in Appendix L.1 — evasion that must leave the cited evidence byte-identical, and hijacking that steers the citation onto a decoy while the verdict stays "fake" — are studied there only in the synthetic sandbox ($n = 16$) and support no claim about a deployed system; no adaptive attack was run against any real model, and we do not claim one (Section K).

Positioning. The individual ingredients – counterfactual testing, matched controls, activation patching, gradient supervision – all exist. The claim of this paper is narrower and, we think, defensible: manipulation forensics admits a held-source counterfactual with an authentic-image identity and no new synthesis trace, and once you have it, faithfulness becomes a measurable quantity with a matched control and a calibration story for the cases where the gold counterfactual is unavailable.

B Study design and evidence scope

This section expands the design rationale and scope of the four contributions in Section 1. Detailed experimental results are reported in the corresponding appendix sections.

Held-source restoration. Paste-back uses a retained authentic source to define replacement pixels, without a learned or classical synthesis step. On an authentic input $x = x_{\mathrm{src}}$, it is the identity for every region and has a zero false-flip rate for a fixed deterministic detector. This algebraic property does not make partial restorations of manipulated images distribution-preserving. Full restoration at $M^\star$ returns the source only when source and target are identical outside $M^\star$. The splice/copy-move corpora satisfy this condition by certificate; MagicBrush does not (Sections C, I and E.13).

Positive controls and separate estimands. The validation protocol evaluates detectors with known dependence mechanisms through the same audit procedure. Its correct-but-non-causal arm detects perfectly and cites the true region while reading a global cue. IoU cannot distinguish that arm from a faithful one; necessity can (Section D). Conversely, necessity alone does not measure completeness of localisation. We therefore report necessity beside IoU, rather than aggregate them into a composite whose ordering closely follows IoU on the evaluated real-data populations (Appendix O).

Area-matched control regions test whether restoration effects are specific to the output region. Suspicion matching narrows the measured sensitivity gap, but its downstream advantage over random controls is unresolved in the evaluated populations. Control coverage and the paired cost of fragmented fallback regions are reported rather than assumed harmless. For synthesized counterfactuals, false-flip calibration and region containment must also be measured: operator effects vary substantially across detectors (Sections C and G).

Released-system audits. FakeShield and FFAA are case studies of released forensic configurations (Sections E and E′). Their checkpoints and prompts are identified, and loader checks precede inference. FakeShield is evaluated with both a post-hoc text grounder and its native MFLM mask. The latter is an output region, not an established narration-conditioned citation. FFAA's single-part experiment maps named anatomy to masks, permitting an own-versus-shuffled-narration comparison. Different corpora, region instruments, readouts and eligible cohorts prevent a direct ranking of the two systems. Necessity and sufficiency are operational intervention effects, not an additive allocation of internal evidence.

Trained-detector validation. The hybrid corpus combines certified local edits with a controlled global cue, making local-evidence and global-cue training arms possible on the same corpus (Section J). The cue is either an analytic field of known strength or a JPEG-75 re-encode. Citation supervision and training budgets are held fixed, but detection calibration and training distributions are not fully matched. Seed/split replication and the exploratory source-exposure and ranking checks in Section P qualify the findings.

Supporting evidence and its limits. Three general-purpose served VLMs provide paired comparisons, not substitutes for the released-system audits (Section F). Activation patching and FaithAlign are evaluated only on locally hosted Qwen2-VL-2B-Instruct; their conclusions do not transfer automatically to other backbones (Section H). The synthetic sandbox provides detector archetypes with known causal structure and a small attack study (Section L.1); it supports no attack or robustness claim about a released forensic system. Section K records these limits, and Table 5 distinguishes planned analyses from choices made after inspecting outcomes.

C Method in full

C.1 Setup

Let f be a forensic VLM with a fixed deterministic manipulation score $p(x) \in [0,1]$, a verdict $y = \mathbb{1}[p(x) \geq \frac{1}{2}]$, and a free-text explanation $e(x)$. Its output region $R(x) \subseteq \Omega$ is a native mask, an emitted box, or a region mapped from $e(x)$ by a stated grounder g; Ω is the pixel lattice. Let $M^\star$ be the ground-truth tamper support and, when available, let x_{src} be the pre-edit authentic source of x.

Definition 1 (Authentic-counterfactual operator)**.** A_R restores region R of x toward the authentic manifold. The *paste-back* operator is

$$A_R^{\mathrm{pb}}(x)[u] \;=\; \begin{cases} x_{\mathrm{src}}[u], & u \in R \\ x[u], & u \notin R, \end{cases} \tag{1}$$

and a *generative* operator fills R from context with an inpainter.

Proposition 2 (Paste-back is the identity on authentic content)**.** *If x is authentic, i.e. $x = x_{\mathrm{src}}$, then $A_R^{\mathrm{pb}}(x) = x$ for every $R \subseteq \Omega$. Consequently the false-flip rate of paste-back on authentic images is exactly 0, and $p(x) - p(A_R^{\mathrm{pb}}(x)) = 0$ identically, for any fixed deterministic detector f.*

Proof. Immediate from Eq. 1: both branches read from the same array. The claim holds pointwise, so it is independent of f, of R, and of any property of the image. □

Proposition 2 is elementary but useful: unconstrained generative filling does not guarantee the same authentic-image identity, and we verify it numerically rather than assume it. The proposition is deliberately narrower than a claim that partial restoration removes all confounding pathways. On a manipulated input, restoring a proper subset can change boundary continuity, patch–context consistency and interacting or redundant cues. Paste-back introduces no *new synthesis trace*; it does not rule out those alternative pathways. The corpus locality certificate and the hard/feathered/recompressed comparison test the conditions we can observe, while the remaining interaction ambiguity is a limitation of the intervention estimate.

Algorithm 1 states the core causal estimator and the point at which each control is drawn. The identity result above proves the authentic-image false-flip rate without scoring a duplicate array; it does *not* make the audit query-free. At $G = 4$, the shipped full ledger has a cold upper bound of 31 audited-model calls per image: the original verdict and explanation, the G^2 occlusion sweep and exact suspicion checks, paste-backs for the cited, random, fragmented, ceiling, and ceiling-control regions, two inpainter checks, and authentic-image baselines. Paste-back itself invokes no generative model. The control is drawn *after* the citation is known and is matched on the model's own occlusion sensitivity; where geometry admits no congruent disjoint translate the area-exact fragmented fallback is used and recorded (Appendix L.2).

C.2 Intervention and localisation quantities

Necessity. Restoring the cited evidence must move the verdict:

$$\mathrm{Nec}(x) \;=\; p(x) - p\big(A_{R(x)}(x)\big)\,. \tag{2}$$

Suspicion and the matched control. The model's own suspicion of a region is its occlusion sensitivity

$$s(R) \;=\; p(x) - p(\mathrm{occ}(x, R))\,, \tag{3}$$

where occ replaces R with a content-free fill. We draw a control R_{ctrl} satisfying

$$R_{\mathrm{ctrl}} \cap R(x) \approx \emptyset, \qquad \big||R_{\mathrm{ctrl}}| - |R(x)|\big| \le \tau_a |R(x)|, \qquad \big|s(R_{\mathrm{ctrl}}) - s(R(x))\big| \le \tau_s\, \sigma_s(x), \tag{4}$$

with $\sigma_s(x)$ the standard deviation of the image's own sensitivity map so that the tolerance means the same thing for a peaked and a flat suspicion landscape. Candidates are translates of the cited region's own footprint, which removes area as a confounder before sensitivity matching begins. Specificity and the causal contrast are then

$$\mathrm{Spec}(x) = p(x) - p(A_{R_{\mathrm{ctrl}}}(x))\,, \qquad \Delta\mathrm{NS}(x) = \mathrm{Nec}(x) - \mathrm{Spec}(x). \tag{5}$$

An explanation is faithful when $\Delta\mathrm{NS} \gg 0$: restoring what it cited matters, and restoring an equally-suspicious region it did not cite does not.

Two readings of Spec coexist and it is worth separating them now, because our experiments support the second more strongly than the first. Read one: Spec detects a detector that is sensitive everywhere, for which necessity alone would be misleading. Read two: Spec is a per-example, region-matched estimate of the counterfactual *operator's* own effect, which ΔNS then subtracts. For authentic inputs, paste-back's effect is zero by Proposition 2; on certified local edits, it is also zero for controls disjoint from $M^\star$. Controls disjoint only from $R(x)$ may still intersect $M^\star$ and have a nonzero effect; under a generative operator it is not zero by construction and the control earns its place — how large it is depends on the detector, and Section G measures it ranging from negligible (big-lama on the global-cue oracle: -0.0000 $[-0.0000, -0.0000]$, 0.0% of authentic images flipped) to sign-inverting. Section L.1 reports both readings against the data.

Algorithm 1: Core paste-back estimator for one image; the full validation ledger is 31 calls at $G = 4$.

Input: image x; held source x_{src}; detector f with explanation head e; grounder g; ground-truth support $M^\star$

Output: necessity Nec, specificity Spec, ΔNS, localisation L_{gt}

$p_0 \leftarrow p(x)$ `// already required by the verdict`
$R \leftarrow g(e(x))$ `// the cited region, or the system's own mask`
$S \leftarrow$ `occlusion_sensitivity`(f, x) `// $G^2$ model evaluations`
$C \leftarrow$ `match`(R, S) `// disjoint, area-matched, suspicion-matched`
if $C = \varnothing$ **then** $C \leftarrow$ `fragmented`(R, S) **and record it**
$x_R \leftarrow A_R^{\text{pb}}(x)$; $x_C \leftarrow A_C^{\text{pb}}(x)$; $x_{M^\star} \leftarrow A_{M^\star}^{\text{pb}}(x)$ `// no synthesis`
$\text{Nec} \leftarrow p_0 - p(x_R)$; $\text{Spec} \leftarrow p_0 - p(x_C)$; $\Delta\text{NS} \leftarrow \text{Nec} - \text{Spec}$
$L_{gt} \leftarrow \text{IoU}(R, M^\star)$ `// reported beside` ΔNS`, never fused`
return $(\text{Nec}, \text{Spec}, \Delta\text{NS}, L_{gt})$ *and the ceiling* $p_0 - p(x_{M^\star})$

Congruence has a geometric limit, and it is not missing at random. A congruent disjoint control does not always exist. On a square canvas a compact citation of area fraction a admits a disjoint translate only if $2\sqrt{a} \leq 1$, i.e. $a \leq 0.25$, and not even then if the citation is centred. The rule therefore fails *exactly* on the rows where the model cited a large region — which is the "cite half the image and you can never be wrong" behaviour the metric exists to punish. Conditioning the audit on a congruent control existing is thus a selection that removes the worst-behaved rows and flatters the models that produce them, and in our data it is not a corner case: it deletes $1 - 0.827$ of the served-model rows and $1 - 0.400$ of the rows produced by the 2B models we train ourselves.

We keep congruence wherever the geometry allows it and replace it where it does not, with an **area-exact fragmented control**. The cited footprint is cut into a $k \times k$ grid of tiles and the tiles are re-placed elsewhere, disjoint from the citation and from each other, at the smallest k that admits a placement. Because the tiles partition the footprint, $|R_{\text{ctrl}}| = |R(x)|$ holds *exactly*, so the "restoring more pixels" confound that congruence was protecting against is still controlled; what is given up is shape. Suspicion matching then proceeds over the fragmented candidates exactly as before.

We do not assume the substitution is free. Three columns are reported throughout: CF under the congruent-only control (with incomplete coverage), under the coverage-complete rule, and — on every row where *both* controls exist — the paired difference between them, which is the direct measurement of what fragmentation costs (Appendix L.2).

Localisation consistency. Faithful is not the same as correct. We report

$$L_{\text{gt}} = \text{IoU}(R(x), M^\star), \qquad L_{\text{attr}} = \text{IoU}(R(x), \text{top}_\alpha\, a_f(x))\,, \tag{6}$$

where a_f is the model's internal attribution of the *verdict*. L_{gt} asks whether the citation is the true tamper; L_{attr} asks whether it is where the model's own verdict computation looks.

The deliverable is the pair, not a product. A candidate aggregation combines the causal and localisation factors into a single composite $\text{CF}(x) = \sigma(\Delta\text{NS}(x)/T) \cdot L(x)$ (Eq. 21, Appendix O). We report its factors separately because on every real population we tested, the composite is a rescaled IoU ($\rho(\text{CF}, \text{IoU}) = +0.990$[+0.985, +0.994] on served rows, +0.961 $[+0.955, +0.967]$ pooled over the control arms, and no value of T in $[0.02, 2]$ drops it below $+0.883$); it is out-ranked by its own causal factor on the positive control; and on the trained-network cell built to decide it, necessity separates the arms and the composite does not. Appendix O reports the composite-score diagnostics and the temperature sweep. Our primary report is

$$\big(\ \text{Nec}(x, R(x))\ \text{ under paste-back},\ \ L_{\text{gt}}(x)\ \big), \tag{7}$$

two numbers reported side by side: a causal fact and a localisation fact, with separate evidence for each, and with the causal one read on the log-odds scale (§C.2) whenever the verdict is saturated.

Necessity on the log-odds scale. Nec as defined in Eq. 2 is a difference of probabilities, and a trained detector sits at the top of the sigmoid: in the MagicBrush global-cue comparison $p(x)$

had mean 0.9985 (min 0.9549) for one arm and 0.9997 (min 0.9985) for another, so the metric's entire dynamic range on 38 of 40 rows was $\sim 10^{-4}$ probability units. Under saturation a probability difference mechanically returns ≈ 0 *whatever* pathway produces the verdict, which makes the design unable to falsify the null it is aimed at. We therefore also report

$$\mathrm{Nec}^{\mathrm{logit}}(x, R) \;=\; \mathrm{logit}\, p(x) \;-\; \mathrm{logit}\, p(A_R(x)), \qquad \mathrm{logit}\, p = \log \tfrac{p}{1-p}, \tag{8}$$

with probabilities clamped to $[\epsilon, 1-\epsilon]$, $\epsilon = 10^{-6}$, and the clamped fraction reported per arm. This is a re-expression of the same stored quantities — no extra query, no retraining — so the two scales are measured on identical rows and can be compared directly. Both are reported everywhere; where they disagree we say so.

The argument for the log-odds scale is *only* the range argument above, and we bound it here so that no later section can be read as claiming more. On the one cell where the *log-odds* re-expression has a seed replication (Section J.5; the JPEG cell of Section J.6 is replicated too, on the raw scale), the log-odds re-expression of the self-cited necessity contrast is the least stable quantity in this paper: per-seed means 0.404, -0.819, -0.358, per-seed p 0.4841, $< 10^{-4}$, 0.2075, sign agreement 67%, 1 of 3 seeds inverting the sign, against a raw-probability region-matched contrast on the identical runs that agrees in sign on every seed and clears $p < 0.05$ on every seed. Read the scale that has range; do not read log-odds as the more reliable scale.

C.3 Two refinements that answer the obvious objections

Is the matched control circular? Occlusion sensitivity is a proxy for flip likelihood, so conditioning on it might mechanically deflate Spec. Three answers. (i) Matching uses a *different intervention* (content-free occlusion) from the evaluation (authentic restoration); they are correlated, not identical. (ii) We measure the residual coupling directly. Over a population of lattice-block regions we fit

$$p(x) - p(A_R(x)) \;=\; \beta_0 + \beta_1 s(R) + \varepsilon, \tag{9}$$

and report the *sensitivity-adjusted necessity*

$$\mathrm{Nec}_{\mathrm{adj}}(x) \;=\; \mathrm{Nec}(x) - \left(\beta_0 + \beta_1 s(R(x))\right), \tag{10}$$

i.e. how much more the cited region matters than a generic region of equal suspicion. This is the continuous generalisation of the matched control: instead of one matched sample it uses the whole candidate population. If matching were doing all the work, $\mathrm{Nec}_{\mathrm{adj}}$ would collapse for every model. In the sandbox study of Appendix L it was instead the sub-score that tracked designed faithfulness most tightly. (iii) The sampler records its achieved matching error on every call, so the audit reports how often matching actually succeeded.

The inpainting confound. Where x_{src} is unavailable, restoration is generative and injects traces. Rather than subtracting a scalar mean false-flip rate, we fit a *region-conditional* model on authentic images – where, by Proposition 2's converse reasoning, any measured effect is attributable to the operator alone:

$$\widehat{\delta}_{\mathrm{op}}(x, R) \;=\; \phi(x, R)^{\top} \theta_{\mathrm{op}}, \qquad \mathrm{Nec}_{\mathrm{cal}} \;=\; \mathrm{Nec} - \widehat{\delta}_{\mathrm{op}}(x, R(x)), \tag{11}$$

with ϕ collecting area, edge density, local noise energy, perimeter ratio and suspicion – the properties that actually control how visible an inpainting artefact will be. Section L.1 reports the effect in the sandbox: uncalibrated necessity under our classical Laplace-fill inpainter is -0.36 (the operator's own trace dominates and the sign inverts); calibrated, it is +0.61, against a paste-back gold value of +0.34. On real served VLMs the raw operator effect is larger still — the inpainters flip 31.3% of *authentic* images on their own (Section L.2) — which is why we report necessity under every operator side by side with its false-flip rate rather than presenting a calibrated number alone.

C.4 Grounding

CF-Score is only as good as $R(x)$, so grounding is an explicitly varied factor rather than a fixed choice. The implementation supports three settings: *native*, using the model's own regions; *post-hoc*, mapping each parsed artifact class to an interpretable image feature and thresholding it at a fixed area budget (an offline stand-in for Grounding-DINO+SAM2); and *human-verified*, using a

stored annotation. The last is an interface, not a human-verified experimental cohort in this paper; the adjudication in Section E.7 uses automated raters. A consensus grounder reports pairwise IoU between grounders as an agreement statistic. Explanations are parsed into artifact phrases against a forensic lexicon with negation handling, and unparseable citations are reported as their own category – never silently merged with "no citation", which would let a model dodge the audit by emitting vague text.

FakeShield's explanation module (DTE-FDM) answers in a fixed prose schema and emits no box, *including* when asked for one in the format the served models obey (0% of 50 audited rows carry a native citation); localisation in that system is delegated to MFLM, a separate SAM-based module, which Section E.6 loads and audits so that the citation leg is measured through the system's own region as well as through ours. The post-hoc grounder remains the only instrument FakeShield's *explanation head* admits; the second released system of Section E′ admits a narration-derived one instead, because a face narration names anatomy. That is a difference in what the two *released formats* make measurable, not a finding about either model — the shuffled-text control does not establish that MFLM's mask is conditioned on what that head wrote (+0.0758 $[-0.0285, +0.1829]$ when handed another image's narrative) — and the comparison in Section E is reported with that asymmetry named. The post-hoc citation is additionally projected onto the densest axis-aligned window of the same area budget, because the counterfactual leg needs an area-exact control placed *elsewhere* and a scattered top-k pixel set has a canvas-sized bounding box (measured: 0.97 here), which admits no such control at all. The projection is exact, area-preserving, and the fraction of the grounder's mass it keeps is recorded per row.

C.5 Mechanistic decoupling

CF-Score is behavioural. The mechanistic question underneath is which pathway produces the citation. Let $z(x) \in \mathbb{R}^{G\times G\times D}$ be the visual tokens. We run the model on $A_{R(x)}(x)$ and splice the original tokens of $R(x)$ back in, measuring how much of the verdict and of the citation distribution is recovered:

$$\mathrm{rec}_v = \frac{\ell_{\mathrm{patch}} - \ell_{\mathrm{cf}}}{\ell_x - \ell_{\mathrm{cf}}}, \qquad \mathrm{rec}_e = 1 - \frac{d_{\mathrm{TV}}(c_{\mathrm{patch}}, c_x)}{d_{\mathrm{TV}}(c_{\mathrm{cf}}, c_x)}. \tag{12}$$

A coupled system recovers both together. When the token channels the verdict reads are identifiable, we go further with channel-resolved path patching: patch only the verdict-relevant channels and measure the citation's response (VCE), then patch only channels the verdict cannot see (NCE), and define the *rationalisation index*

$$\mathrm{RI} \;=\; \frac{\mathrm{NCE}}{\mathrm{VCE} + \mathrm{NCE}}. \tag{13}$$

$\mathrm{RI} \to 1$ means the citation is driven by information the verdict never consults: rationalisation in the strict sense, not merely a poorly chosen region.

C.6 FaithAlign

Four losses plus an anti-gaming regulariser:

$$\mathcal{L} = \lambda_v \underbrace{\mathcal{L}_{\mathrm{NLL}}}_{\text{verdict}} + \lambda_g \underbrace{\mathcal{L}_{\mathrm{ground}}}_{\text{cite } M^\star} + \lambda_c \underbrace{\mathcal{L}_{\mathrm{cf}}}_{\text{counterfactual}} + \lambda_a \underbrace{\mathcal{L}_{\mathrm{attr}}}_{\text{attribution}} + \lambda_s \underbrace{\mathcal{L}_{\mathrm{size}}}_{\text{region size}}, \tag{14}$$

with

$$\mathcal{L}_{\mathrm{cf}} = \big[m_c - (p(x) - p(A_{M^\star}(x)))\big]_+ + \big[\,|p(x) - p(A_{R_{\mathrm{ctrl}}}(x))| - \epsilon_c\big]_+, \tag{15}$$

$$\mathcal{L}_{\mathrm{attr}} = \Big[m_a - \big(\bar{a}_{\mathrm{cited}} - \bar{a}_{\mathrm{ctrl}}\big)\Big]_+, \qquad a_p = \Big|\partial \ell / \partial z_p\Big|, \tag{16}$$

where $\bar{a}$ denotes mean normalised attribution mass over a patch set. $\mathcal{L}_{\mathrm{attr}}$ uses exact second-order autograd through the verdict logit to optimise attribution alignment. This gradient constraint alone does not establish causal faithfulness. The real-VLM experiments train neither $\mathcal{L}_{\mathrm{attr}}$ nor $\mathcal{L}_{\mathrm{size}}$; their objective coverage and the two evaluated forms of $\mathcal{L}_{\mathrm{cf}}$ are stated in Section H.

On differentiability. Paste-back and control sampling are *not* differentiable and we do not pretend otherwise. They are environment steps executed outside the graph and cached, exactly like a data augmentation; the gradient flows through the model's evaluation of the resulting images. This suffices because Eqs. 15–16 are hinges on p and on ℓ, not on the counterfactual construction. Controls are re-drawn against the current model every few epochs, so the model cannot satisfy the control hinge by relocating its suspicion.

C.7 THREAT MODELS

A – explanation-preserving evasion. Find the smallest δ with $p(x+\delta) < \frac{1}{2}$ subject to $\delta|_{R(x)} = 0$. If the citation genuinely drives the verdict, the attacker must modify pixels they are forbidden to touch. **B – explanation hijack.** Perturb a genuine forgery so that the verdict stays "fake" but the emitted region moves onto a decoy, so region-based review inspects clean pixels. Both are decided end-to-end: success is measured by re-running the model *and the grounder* on the perturbed image, never by a surrogate, and evasion additionally verifies that the cited pixels are byte-identical. Attacks use SPSA so the identical attacker applies to non-differentiable stand-ins and to a differentiable defended model; the executed attacks are sandbox-only. No white-box adaptive attack is evaluated on a real VLM.

Rather than proving a decoupled-heads theorem – which would assume the conclusion – we fit an *operational* certificate. On a training split we fit a lower-quantile regression $\widehat{\epsilon}^{\star}(x) = a + b\,\Delta\mathrm{NS}(x)$ so that a target fraction of attack costs lies above the line, then validate coverage, slack and rank correlation on held-out examples (Section L.1). A bound that predicts held-out attack cost is a result; a bound that does not is reported as a negative one.

D THE VALIDATION PROTOCOL IN FULL: A POSITIVE CONTROL WITH A DISCRIMINATING ARM

How to read this section. We report necessity and localisation separately (Eq. 7) on the same 400 rows. Supplementary tables also report the candidate composite CF to test whether aggregation adds information:

oracle arm (faithfulness known by construction)	necessity at $R(x)$	IoU
faithful ($\alpha = 1$)	0.9914 $[0.9912, 0.9916]$	0.7461 $[0.7315, 0.7592]$
global cue (correct, non-causal)	**0.0000** $[0.0000, 0.0000]$	**0.7461** $[0.7315, 0.7592]$
decoy (unfaithful citation)	0.0002 $[0.0001, 0.0004]$	0.0018 $[0.0010, 0.0030]$
half-evidence	0.9916 $[0.9914, 0.9918]$	0.4174 $[0.4096, 0.4241]$

The global-cue row is the whole argument for reporting a causal quantity at all: its IoU is *bit-identical* to the faithful arm's and its necessity is zero, so an IoU-only report scores a detector whose verdict comes from a canvas-wide statistic exactly as it scores a faithful one. The half-evidence row is the argument for keeping IoU: its necessity is as high as the faithful arm's (0.9916 $[0.9914, 0.9918]$ against 0.9914 $[0.9912, 0.9916]$) because what it cites really does drive the verdict — it just cites half of what it should, which only IoU sees (0.4174 $[0.4096, 0.4241]$ against 0.7461 $[0.7315, 0.7592]$). Neither number alone orders these arms correctly; the pair does, and the product of the two (Appendix O) adds nothing over the pair.

Every served model in Section L.2 scores near its own chance reference. Read on its own, that observation is worth nothing, because it is equally consistent with two incompatible worlds: the models really are unfaithful, or *the metric is broken and nothing can score well on it*. A metric on which no object can score high is not a measurement, it is a constant, and no amount of additional served models would tell the two worlds apart. The only thing that can is a **positive control**: detectors whose faithfulness is known by construction, audited by the identical code path, on the identical real images.

The construction. On the same real MagicBrush triples $(x_{\mathrm{src}}, x, M^{\star})$ we define an *oracle* detector. Let $W = M^{\star} \cap \{u : \|x_u - x_{\mathrm{src},u}\|_1/3 > \varepsilon\}$ be the pixels that genuinely changed, and for an

arbitrary query image z let

$$m(z) = \frac{1}{|W|} \sum_{u \in W} \exp\Big(- \frac{\|z_u - x_u\|_1}{3h} \Big), \qquad p(z) = \sigma\big(g\,(m(z) - \tau)\big). \tag{17}$$

The detector asks whether the edited content is still present in the edited region. Four properties follow *by construction*, not by measurement: $p(x) \approx 1$ and $p(x_{\mathrm{src}}) \approx 0$, so detection is perfect and the detection confound of Section L.2 cannot operate; paste-back over any region containing W destroys the evidence, so necessity is large; paste-back over any region disjoint from W leaves m *exactly* unchanged, so specificity is exactly zero; and occluding a lattice cell reduces m in proportion to $|\mathrm{cell} \cap W|$, so the detector presents a genuine, smooth suspicion landscape and the control matcher faces a real problem rather than a flat field.

Crucially, **every arm of the citation family shares Equation 17**. Those arms differ only in what the explanation *cites*, so verdict accuracy is held exactly fixed — we verify in code that the two extreme arms return bit-identical probabilities on every image either of them is ever shown. A gap in CF-Score between them therefore cannot be a detection gap. (Two further families, introduced below, deliberately vary the *mechanism* instead while holding the citation fixed; they match detection in accuracy — AUC 1.000 and flag rates $1.000/0.000$ — rather than to the bit, and we say so wherever they are used.) The citation is a box congruent to the ground-truth box, displaced toward a disjoint decoy by $(1 - \alpha)\, d_{\mathrm{sep}}$, where d_{sep} is the smallest displacement that separates them: $\alpha{=}1$ is the **oracle-faithful** arm, $\alpha{=}0$ the nearest miss, and a separate **oracle-unfaithful** arm cites an unrelated congruent region elsewhere in the image. The box *size* is constant along the sweep, so a rising CF-Score cannot be explained by "a larger citation restores more pixels".

The arm that makes the control decidable. The construction above varies *where the detector points* while holding *how it decides* fixed. That answers one question — does CF-Score respond to localisation — but does not isolate verdict dependence: in the faithful-versus-decoy contrast the two arms differ in IoU (0.746 versus 0.002), so **plain IoU reproduces the headline separation exactly** and the control cannot show that the counterfactual machinery is doing any work. We therefore add the one arm that can decide it. The **global-cue** oracle emits the *ground-truth box* as its citation — so its IoU is identical to the faithful arm's, to the bit, by construction — while its verdict is

$$s(z) = \frac{1}{|\bar{B}|} \sum_{u \notin B} \frac{\|z_u - x_{\mathrm{src},u}\|_1}{3}, \qquad p(z) = \sigma\big(g\,[\log_{10} s(z) - \log_{10} \tau]\big), \tag{18}$$

where B is the region it cites. s is the generator's *global re-encode fingerprint*: DALL-E 2 rewrites the whole canvas, so on this corpus the mean absolute target-source difference *outside* the edit has median 0.0056 and is never exactly zero ($n = 108$ eligible triples, minimum 5×10^{-5}). This is not a strawman: it is the cue we measure `qwen2.5vl:7b` using in Section L.2, where restoring the true mask recovers only 0.124 of its 0.453 total separation. Because s reads pixels outside B and paste-back rewrites pixels inside B, **necessity is exactly zero by algebra**, and τ is calibrated once below the corpus minimum so detection stays perfect. This is the "unfaithful-but-correct" cell we call the dangerous one, and it is the only arm that can falsify "CF-Score is IoU with extra steps".

Three further arms probe things a single scalar would confuse. The **half-evidence** oracle reads, and honestly cites, only half the manipulation: faithful but incomplete. The **noisy** oracle adds 2 nats of logit noise to the same verdict: faithful but poorly calibrated. The **blind anomaly detector** is the only unprivileged arm — it never sees the source or the mask, scores the mean high-pass noise anomaly of a contiguous block and cites exactly that block, so its verdict is *literally computed from the pixels it cites*.

Design of Table 6. Detectors of known faithfulness on 80 real MagicBrush triples $\times$ 5 seeds (10 arms, 4000 audited rows; 6000 including the mechanism sweep of Table 7), scored through the identical audit used for the served models.

Reading Table 6. `AUC` is 1.000 for every arm except `blind_anomaly`, the one unprivileged detector. The faithful-to-decoy citation sweep shares one scoring function and varies only the cited region. The half-evidence, noisy and blind-anomaly arms probe additional properties. The `oracle_global_cue` arm varies the mechanism, holding the citation at the ground-truth box and matching detection accuracy (AUC 1.000, flag rates $1.000/0.000$, identical to the faithful arm)

Table 6: Positive control: detectors of known faithfulness, same audit.

Detector	n	AUC	IoU only	Nec	dNS	CF	sd across seeds
ORACLE-FAITHFUL (alpha=1)	400	1.000	0.746	0.991	0.991	0.729 [0.713, 0.743]	0.0111
oracle faithful, noisy verdict	400	1.000	0.746	0.960	0.960	0.727 [0.710, 0.740]	0.0112
oracle half-evidence	400	1.000	0.417	0.992	0.992	0.410 [0.402, 0.416]	0.0042
partial alpha=0.75	400	1.000	0.419	0.852	0.848	0.405 [0.395, 0.414]	0.0061
partial alpha=0.50	400	1.000	0.179	0.191	0.174	0.121 [0.114, 0.127]	0.0044
partial alpha=0.25	400	1.000	0.049	0.007	−0.000	0.023 [0.022, 0.025]	0.0012
nearest miss alpha=0	400	1.000	0.002	0.000	−0.061	0.000 [0.000, 0.001]	0.0003
ORACLE-UNFAITHFUL (decoy)	400	1.000	0.002	0.000	−0.032	0.000 [0.000, 0.001]	0.0003
GLOBAL-CUE (true box, non-causal verdict)	400	1.000	0.746	0.000	−0.001	0.370 [0.362, 0.378]	0.0056
blind anomaly detector	400	0.332	0.063	−0.068	−0.055	0.027 [0.023, 0.032]	0.0020

while reading its verdict from outside the region it cites. The *IoU only* column is the competing metric, printed so a reader can check which contrasts it reproduces: it reproduces every one of them except the global-cue row, where it is identical to the faithful arm's by construction. Brackets are 95% bootstrap CIs; the last column is the standard deviation of the arm's mean CF-Score across the 5 seeds. A seed selects the subsample of the eligible pool and is threaded into the audit config (it draws the random control region and breaks decoy-placement ties), so the seeds are genuinely different experiments rather than one experiment repeated — a seed that did not reach the audit config would give an across-seed spread of identically zero, which is the failure this threading prevents.

Reading Figure 4. Left: CF-Score for every arm. Blue bars are detectors of known faithfulness, red bars are the served VLMs of Section L.2 on the same corpus and the same operator, and the black tick is each served model's unfaithful-citation null. Right: CF-Score, $\Delta\mathrm{NS}$ and IoU as the citation is swept from an unrelated region to the true evidence.

Result: the metric separates known-faithful from known-unfaithful, on every image where both are defined. The oracle-faithful arm scores $\mathrm{CF} = 0.729[0.713, 0.743]$; the oracle-unfaithful arm, *with bit-identical verdicts on every image*, scores $0.000[0.000, 0.001]$. The paired difference is $0.729[0.697, 0.757]$, on 356 audited pairs drawn from 96 distinct images, with a cluster-level sign-flip permutation $p < 10^{-4}$; the faithful arm wins on every one of them, and the across-seed standard deviation of its mean is 0.011 against a gap of 0.73. *All resampling here is clustered on the image.* The five seeds each draw 80 triples from an eligible pool of 108, so the 400 rows per arm contain only 96–108 distinct images and treating them as independent narrows every interval by about $\sqrt{3.7} \approx 1.9$; the row-level interval was $[0.713, 0.743]$ and the correct one is wider. We also report the permutation p as $< 1/(B{+}1)$ rather than as the measured value it is not. CF-Score is therefore not a constant, and the near-chance served-model numbers are a statement about those models rather than about the instrument.

Result: the causal factor does work that localisation cannot — the test that could have gone the other way. The global-cue oracle detects perfectly (AUC 1.000, flag rate 1.000 on edits and 0.000 on sources, the same accuracy as the faithful arm), cites the true box (IoU 0.746, identical to the faithful arm's 0.746), and has necessity 0.0000 and $\Delta\mathrm{NS} = -0.0011$. It scores $\mathrm{CF} = 0.370[0.362, 0.378]$ against the faithful arm's $0.729[0.713, 0.743]$:

paired ΔCF (faithful − global cue)	0.358 $[0.343, 0.372]$, $p < 10^{-4}$, 356 pairs over 96 images, faithful wins 100%
paired ΔIoU (same pairs)	0.000 $[0.000, 0.000]$ — zero by construction

A metric that reported localisation alone would score these two detectors *identically*; CF-Score separates them by half its measured range, and the separation is carried entirely by $\sigma(\Delta\mathrm{NS}/T)$, which is ≈ 1 for one arm and ≈ 0.5 for the other. The redundancy diagnostics complement this controlled comparison: pooled over all arms, $\rho(\mathrm{CF}, \mathrm{IoU}) = +0.961$ $[+0.955, +0.967]$ and $\rho(\mathrm{CF}, \Delta\mathrm{NS}) = +0.723$ $[+0.708, +0.738]$, and across arm means $\rho(\mathrm{CF}, \mathrm{IoU}) = +0.84$ — high despite the discriminating global-cue contrast. The honest summary is that IoU carries most of the *variance* of CF-Score and cannot carry the one contrast that matters.

Result: a second monotone sweep, on the axis a localisation metric cannot see. The global-cue arm invites one objection: we defined its statistic to be invariant to paste-back, so of course its

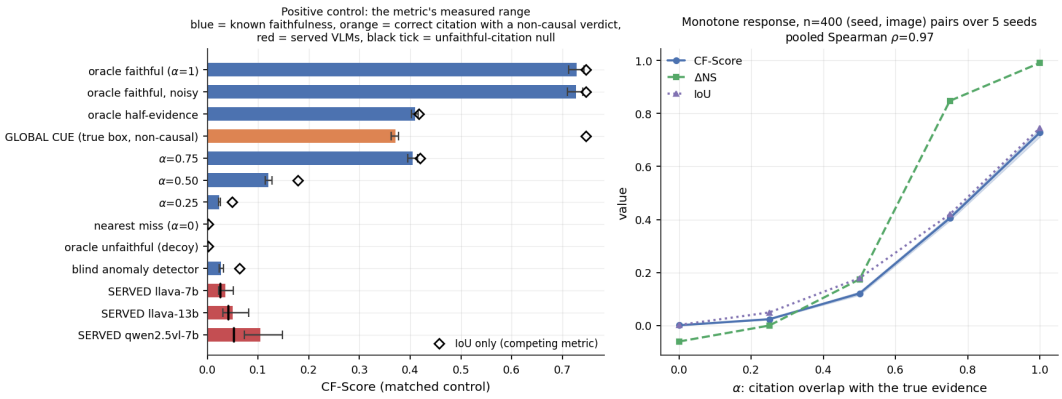


Figure 4: **The metric's measured range, and the served models inside it.**

necessity is zero. The answer is that the arm is one endpoint of a family. `mixed_cue`(β) takes the logit to be β times the local edit-template term plus $(1-\beta)$ times the global-cue term, and **cites the ground-truth box at every** β — so localisation is held identical *within every image* (measured within-image IoU spread: $0.00e+00$) and detection AUC is 1.000 throughout, while β weights the local contribution to the constructed logit. It is not an additive share of the probability-scale intervention effect. Sweeping β over 0.00, 0.25, 0.50, 0.75, 1.00:

IoU only (competing metric)	0.7461 at every β	range $0.00e+00$
necessity	0.000, 0.083, 0.524, 0.927, 0.991	
CF-Score	**0.370, 0.431, 0.659, 0.725, 0.729**	range 0.358

CF-Score moves across *half its entire measured range* while the localisation metric is a flat line, with a per-image Spearman $\rho(\mathrm{CF}, \beta) = 1.000$ on 100% of the 356 images with a complete curve. The endpoints reproduce the two named arms exactly (β=0 is `oracle_global_cue` at 0.370, β=1 is the faithful oracle at 0.729), which is an internal consistency check on the whole construction. This is the sweep the alpha family cannot provide, and it is the direct measurement that CF-Score's causal factor tracks the causal structure with geometry held fixed.

Reading Table 7. Every arm cites the ground-truth box, so *IoU only* is identical within every image and detection AUC is 1.000 throughout; β is the fraction of the verdict logit contributed by local evidence rather than by a canvas-wide generator fingerprint. A metric that reported localisation alone is constant down this table.

Reading Figure 5. Citation is the true box at every point, so the competing metric (IoU only) is flat by construction; CF-Score traverses half its measured range as the verdict's causal support moves from a canvas-wide cue onto the manipulated pixels.

Result: the response is monotone in citation faithfulness. Sweeping α from 0 to 1 moves CF-Score through 0.000, 0.023, 0.121, 0.405, 0.729 — strictly increasing at every step, with a pooled Spearman $\rho = 0.97$ against α and a positive within-image rank correlation on 100% of the 260 (of 400) images where all five points of the curve are defined under the congruent control. The other 140 are dropped from the monotonicity statistic, not from the table: at small α the citation can land where no congruent disjoint control fits. Under the coverage-complete leg of Section L.2 every arm here has coverage 1.000 and the known-truth separation is unchanged — faithful minus decoy 0.7314 $[0.7033, 0.7568]$ ($p = 0.0001$, $n = 400$) on all 400 pairs, against 0.729 $[0.697, 0.757]$ on the 356 the congruent control could serve — which is the check that the fragmented substitution does not disturb a case whose answer we know. Because the citation area is held constant along the sweep, this is a response to *where* the explanation points and nothing else.

Result: the two factors of CF-Score are not redundant. The half-evidence oracle is maximally faithful ($\Delta\mathrm{NS} = 0.992$, the same as the full oracle) yet reaches only $\mathrm{CF} = 0.410[0.402, 0.416]$, because it correctly cites half of a manipulation and IoU falls to 0.417. This is the "honest but

Table 7: Mechanism sweep: citation fixed, causal structure varied.

beta (fraction of the verdict from LOCAL evidence)	n	AUC	IoU only	Nec	dNS	CF
0.00	400	1.000	0.7461	0.000	−0.001	0.370 [0.362, 0.378]
0.25	400	1.000	0.7461	0.083	0.082	0.431 [0.421, 0.440]
0.50	400	1.000	0.7461	0.524	0.522	0.659 [0.645, 0.672]
0.75	400	1.000	0.7461	0.927	0.927	0.725 [0.709, 0.739]
1.00	400	1.000	0.7461	0.991	0.991	0.729 [0.713, 0.743]

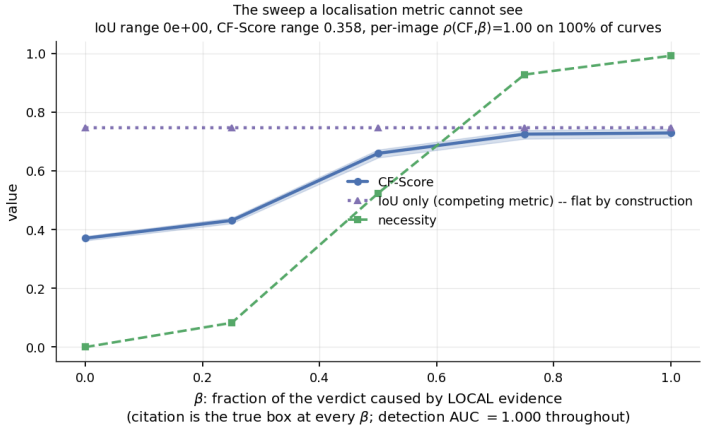


Figure 5: The sweep a localisation metric cannot see.

incomplete" cell, and a scalar that fused faithfulness with correctness would report it as a partial failure of faithfulness, which it is not. The noisy oracle scores 0.727 $[0.710, 0.740]$: CF-Score does not simply track how deterministic a verdict is.

Evidential necessity is not computational dependence. The blind anomaly detector computes its verdict *from the pixels it cites*, which is the strongest imaginable form of mechanistic faithfulness — and it scores CF $= 0.027[0.023, 0.032]$ — an order of magnitude above the unfaithful oracle's 0.000 $[0.000, 0.001]$, and more than an order of magnitude *below* the faithful arm's 0.729 $[0.713, 0.743]$, so it sits at the floor of the scale rather than exactly on the unfaithful arm. The reason is instructive and is a property of the definition rather than a bug: paste-back restores *authentic* pixels, so on a region that was never edited it is the identity map and cannot move any verdict. CF-Score therefore measures *evidential* necessity — does the cited region carry the manipulation — and not *computational* dependence — is the verdict a function of those pixels. For forensics the former is the quantity that matters, since an explanation citing unedited pixels is not useful to an analyst however mechanistically faithful it is. The two concepts are distinct, and a reader auditing a detector whose evidence is global (Section L.2) must read Nec against the restore-the-true-mask ceiling for exactly this reason.

The closest prior work, on the load-bearing evidence — and it makes the decisive contrast too. Were EDCT's CCS (Ding et al., 2025) computed only in the synthetic sandbox, i.e. in none of the three bodies of evidence this paper leads with, the only comparator CF-Score had on the evidence that matters would be localisation IoU. EDCT's *answer leg* needs nothing we have not already bought — the verdict on x and the verdict on $A_{R(x)}(x)$ — so we recompute it on the identical rows, at zero additional forward passes, and report it here. Its explanation leg is not computable on these arms and we do not fake it: an oracle's citation is bound to the sample rather than read off the image, so re-explaining a counterfactual returns the same string by construction. Every CCS number below is therefore the answer leg alone, with paste-back as the operator — i.e. EDCT's criterion given *our* counterfactual.

The comparison limits the score-specific novelty. On the global-cue arm, CCS's answer leg fires on 0.000 of rows against 1.000 for the faithful arm — a paired gap of 1.000 $[1.000, 1.000]$ ($p < 10^{-4}$, 400 pairs). **CCS-with-paste-back separates the correct-but-non-causal detector exactly as well as CF-Score does.** The composite's advantage on that contrast is therefore not over the closest prior

work; it is over prior work *using a synthesised counterfactual*, which is an argument for the operator, not for the score. That is the honest reading and we adopt it: the operator and the validation protocol are the contribution, and Section K states it in those words.

What CCS does not buy is resolution, and that is measurable rather than rhetorical. CCS is a threshold crossing, so it saturates: across the citation sweep it takes the values 0.000, 0.000, 0.090, 0.960, 1.000 (saturated at 0 or 1 on 60% of the sweep) where CF-Score takes 0.001, 0.024, 0.112, 0.403, 0.732, and across the mechanism sweep 0.000, 0.010, 0.703, 1.000, 1.000 against 0.372, 0.433, 0.663, 0.728, 0.732. And it has one blind spot with a known answer: the half-evidence oracle is faithful but *incomplete*, and CCS scores it 1.000 — identical to the fully faithful arm's 1.000 — while CF-Score separates them (0.7322 against 0.4097) because the localisation factor sees the missing half. On the served rows CCS fires on 0.120 of `qwen2.5vl:7b`'s rows, and on our own trained arms it rises from 0.000 (untrained) to 0.236 (`faithalign`), so it does register the training effect — more coarsely than Nec, and in the same direction.

What this licenses. The positive control converts every CF-Score in this paper from an uncalibrated number into a position on a scale with a measured ceiling (0.729) and a measured floor (0.000) obtained on the same images, with the same operator, at the same citation areas. Section L.2 places the served models on that scale.

E THE PRIMARY AUDIT IN FULL: A RELEASED EXPLAINABLE FORENSIC VLM

We evaluate a released forensic system on the same images and audit procedure as the served-model comparison. FakeShield v1-22B (DTE-FDM) (`zhipeixu/fakeshield-v1-22b`, Apache-2.0, Xu et al., 2025b) is audited by the *identical* instrument — the same `audit_sample`, the same `AuditConfig`, the same corpus, and the same 50 `sample_id`s as the served-model contrast of Section F, so every number below is paired image-by-image with every number there. This is the paper's audit population; the served chat models are the control group.

What the checkpoint actually is. The published name is misleading in a way worth stating, because a reader comparing parameter counts will otherwise be misled by our table too. The released repository holds three parts: DTE-FDM, the detection-and-explanation model this audit runs; MFLM, a separate GLaMM/SAM localiser; and a small ResNet-50 domain-tag generator that supplies the prompt prefix the released inference code prepends (we use it, so the model is prompted the way its authors prompt it). "v1-22B" counts DTE-FDM and MFLM together. DTE-FDM alone declares `_name_or_path="liuhaotian/llava-v1.5-13b"` and holds 13.351 B parameters — it is a **LLaVA-1.5-13B-class** model, and the right same-scale comparison in the tables below is therefore `llava:13b`, not the 7B models. The weights are public, ungated and Apache-2.0; the checkpoint we loaded is 6 shards, 758 tensors, 27.309 GB on disk, exactly matching the `total_size` in its own `model.safetensors.index.json`.

What was run, and the evidence that it was really run. The checkpoint's own `config.json` names an architecture class that exists only in the original LLaVA repository: `architectures=["LlavaLlamaForCausalLM"]`. We remap it onto `transformers`' native `LlavaForConditionalGeneration`: a bijective rename of 758 tensors, no reshapes, no splits. Fluency is not evidence that this worked — a randomly-initialised projector produces confident forensic prose — so the loader refuses to return a model unless the checkpoint and parameter key sets are equal, every shape matches, nothing is left on `meta`, and ten fingerprinted tensors (all four projector tensors included) are **bitwise identical** to the safetensors bytes. The decisive control is behavioural: replacing `mm_projector` with shape-and-scale-matched random weights collapses the output to an empty string or degenerate looping, content-word Jaccard 0.0 against the real run. Decoding is greedy; the released demo samples at $T = 0.2$, which would put sampling noise inside every counterfactual difference. Budget: 1482 local forward passes and 50 generations, 0.68 s per verdict and 69 s per sample on MPS, 57 minutes of wall clock, zero network requests.

The explanation head emits no citation. DTE-FDM answers in a fixed prose schema and emits no box — 0% of rows carry a native citation — *including* when asked for one in the format all

three served models obey (Section C.4). Localisation in that system is delegated to MFLM, a separate SAM-based module that is not part of the explanation — which we load, gate and audit in its own right, so the citation leg is reported through *both* instruments (Section E.6). The post-hoc grounder is the only instrument FakeShield's *explanation head* admits, and that asymmetry — FakeShield grounded post-hoc, the served models by their own boxes — is a genuine confound on every *between-model* citation comparison below. Two things bound it: each model is compared against *its own* area-matched random-citation null, and the causal quantities are also reported on regions that involve no grounder at all.

E.1 Detection first

Faithfulness is only defined where a detector detects, so the verdict comes before any faithfulness number.

Model	n	flag edit	flag auth	margin	AUC	win rate	always-fake
fakeshield-v1-22b/DTE-FDM	50	0.90	0.82	+0.08	0.587 [0.520, 0.663]	0.70	False
llava:13b	50	0.72	0.66	+0.06	0.567 [0.523, 0.620]	0.66	False
llava:7b	50	1.00	1.00	+0.00	0.627 [0.574, 0.687]	0.76	True
qwen2.5vl:7b	50	0.80	0.20	+0.60	0.868 [0.800, 0.930]	0.92	False

FakeShield reaches paired AUC 0.587 $[0.520, 0.663]$, ranking the edit above its own source photograph on 70% of pairs. The interval excludes chance, so the signal is real — and it is *small*: the best served model, `qwen2.5vl:7b`, reaches 0.868 $[0.800, 0.930]$ on the same 50 pairs, a gap of -0.281 $[-0.358, -0.198]$ AUC against the purpose-trained system. Thresholded, FakeShield flags 0.90 of edited images and 0.82 of the authentic photographs they were made from, a usable margin of only +0.08. This is the first result of the section and it is not the one we expected: the released forensic model is a substantially weaker detector on this corpus than a general-purpose VLM asked a forensics question.

E.2 Local sensitivity and an unresolved citation effect on MagicBrush

Model	n	Nec at M*	Spec at area-matched ctrl	dNS ceiling
fakeshield-v1-22b/DTE-FDM	50	0.0527 [0.0127, 0.0901]	−0.0122 [−0.0197, −0.0068]	0.0649 [0.0253, 0.1040]
llava:13b	50	0.0353 [0.0064, 0.0710]	0.0007 [−0.0082, 0.0065]	0.0347 [0.0052, 0.0710]
llava:7b	50	0.0124 [0.0049, 0.0205]	−0.0000 [−0.0014, 0.0018]	0.0125 [0.0049, 0.0203]
qwen2.5vl:7b	50	0.1236 [0.0762, 0.1830]	0.0257 [0.0031, 0.0694]	0.0979 [0.0348, 0.1635]

Reading Figure 6. The *Nec ceiling* bar restores the true edit mask; the *Nec* bar restores the region the explanation cites; the two *Spec* bars restore control regions the explanation did not cite. For FakeShield the ceiling bar is positive with an interval clear of zero — the verdict does respond to the edited pixels — while its contrasts with both control bars have intervals containing zero. The contrast between restoring what the model points at and an equally suspicious region it did not point at is therefore unresolved. Bars are means with BCa bootstrap intervals over paired images.

The grounder-free leg restores the *true* edit mask and compares it against an area-exact control region elsewhere — identical regions for every model, no citation involved. FakeShield's verdict does move: Nec at $M^\star$ is 0.0527 $[0.0127, 0.0901]$ and the controlled difference is 0.0649 $[0.0253, 0.1040]$, an interval excluding zero. So there *is* local evidence in these images that this model's verdict uses, and it is more of it than either LLaVA carries (`qwen2.5vl:7b` remains ahead at 0.0979 $[0.0348, 0.1635]$).

For the citation leg, Restoring the region the explanation points at gives $\Delta\mathrm{NS} = 0.000$ $[-0.024, 0.027]$ on all 50 rows — centred on zero, with restoring the cited region and restoring an equally-suspicious control region unresolved; this does not establish equivalence. CF-Score is 0.056 $[0.036, 0.098]$ against this model's own area-matched citation null of 0.046 — an excess of +0.009, and the interval does **not** clear the null (no). Against the measured floor and ceiling of the oracle positive control (on the same images), where a faithful-by-construction detector scores 0.748 $[0.707, 0.776]$ and a decoy-citing one scores 0.001 $[0.000, 0.004]$, FakeShield sits at 0.07 of the way from floor to ceiling.

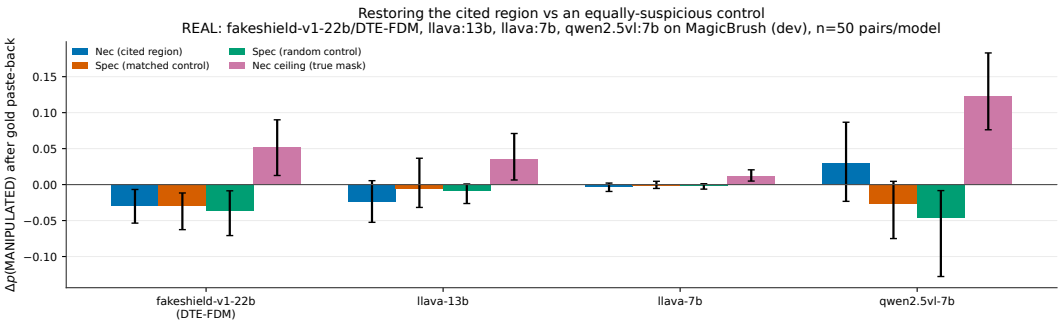


Figure 6: The conjunction this section rests on, for FakeShield v1-22B (DTE-FDM) and the three served chat models on the identical 50 pairs.

An unresolved effect is not evidence of absence. Two results constrain its interpretation. *(a) Power.* On this corpus the smallest citation effect the design could have detected is 0.0374 — 80% of this model's *entire* detection margin 0.0466 $[0.0103, 0.0867]$ — so an interval containing zero excludes very little (Section E.8). *(b) Replication.* Run on a corpus whose evidence is local by certificate, the identical instrument finds a citation effect that *does* exclude zero (0.0233 $[0.0050, 0.0563]$), which limits generalisation of the MagicBrush result (Section E.3). The magnitude is corpus-dependent: on the corpus with the resolution to measure it, restoring the cited region has a normalized intervention effect of 8% $[1\%, 17\%]$ — relative to restoring $M^\star$, and not a share of the evidence the model uses. The conjunction the section rests on is therefore not "verdict causal, citation not" but "verdict causal, citation causal to roughly a twelfth of that degree, and unresolved against a suspicion-matched control on the corpus whose minimum detectable effect is 80% of the detection margin."

E.3 Audit on a certified local-manipulation corpus

Everything above runs on MagicBrush, and Section I of this same paper says of MagicBrush that most of the verdict lives in pixels no local intervention reaches. On that corpus "the citation has no local effect" and "nothing here has much local effect" are not separated, so the dissociation requires a higher-resolution test. We therefore ran the *identical* instrument — same checkpoint, same `AuditConfig`, same 31 forward passes per sample, same $n = 50$, same analysis file, paired against the same served model's `sample_ids` — on the splice/copy-move corpus, whose certificate is that the outside-mask difference is **exactly** 0 on every row. Budget: 50 samples, 1461 local forward passes and 50 generations, 2992 s of MPS wall clock, zero network requests.

Quantity	MagicBrush	Splice, hard	Splice, hard, n=150	Splice, feather
n	50	50	150	50
AUC	0.587 [0.520, 0.663]	0.981 [0.955, 0.999]	0.975 [0.958, 0.989]	0.952 [0.912, 0.986]
margin $p(x) - p(x_{\text{src}})$	0.047 [0.010, 0.087]	0.286 [0.247, 0.325]	0.281 [0.260, 0.305]	0.263 [0.222, 0.304]
Nec at $M^\star$	0.053 [0.013, 0.090]	0.286 [0.247, 0.325]	0.281 [0.260, 0.305]	0.263 [0.222, 0.304]
ΔNS ceiling	0.0649 [0.0253, 0.1040]	0.2856 [0.2468, 0.3247]	0.2813 [0.2603, 0.3053]	0.2629 [0.2222, 0.3041]
ΔNS citation, mean	0.0001 [−0.0242, 0.0275]	0.0233 [0.0050, 0.0563]	0.0198 [0.0104, 0.0331]	0.0041 [−0.0164, 0.0232]
ΔNS citation, median	−0.0027 [−0.0150, +0.0033]	+0.0025 [+0.0000, +0.0094]	+0.0010 [+0.0000, +0.0065]	+0.0000 [−0.0012, +0.0035]
ΔNS citation, L2O mean	−0.0118	+0.0090	+0.0147	+0.0052
sign test p (+/−)	0.480 (22/28)	0.020 (29/13)	0.003 (79/45)	0.551 (25/20)
CF-Score	0.056 [0.036, 0.098]	0.070 [0.050, 0.096]	0.074 [0.061, 0.089]	0.068 [0.049, 0.092]
own null	0.046	0.046	0.046	0.046

Reading Figure 7. Two things change and both matter. The *Nec ceiling* bar is equal to the detection margin by identity — it *equals* the whole detection margin, because paste-back over $M^\star$ returns the authentic image bit-for-bit — and its control bar is exactly 0, since paste-back over a region outside the mask rewrites no pixel at all. The cited-region bar, whose control contrasts included zero on MagicBrush, is small but separated from them in the BCa mean intervals. The interpretation is therefore corpus-dependent. Bars are means with BCa bootstrap intervals over paired images.

A small, estimator-sensitive citation effect. On the hard-paste splice corpus, the cited-region mean is $\Delta\text{NS} = 0.0233[0.0050, 0.0563]$, with a BCa interval excluding zero, and the composite 0.070 $[0.050, 0.096]$ it does clear the area-matched random-citation null (0.046). This differs from

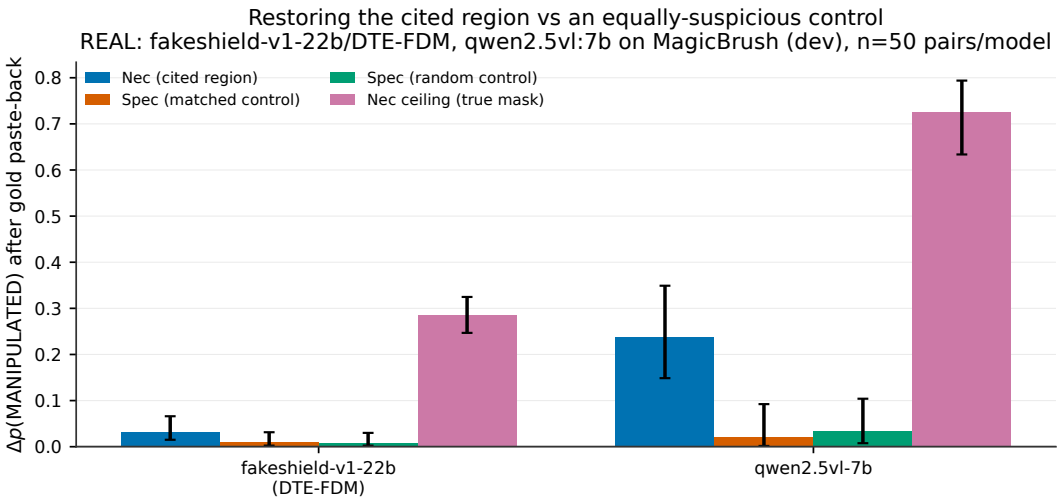


Figure 7: The same four bars as Figure 6, on the splice/copy-move corpus where the outside-mask difference is exactly 0 by certificate.

the unresolved MagicBrush estimate. The following robustness checks qualify the result: the mean is sensitive to two rows, the paired permutation test is not significant at $n = 50$, and the feathered construction does not resolve the same effect.

Robustness across estimators. We report the same paired estimators used for the white-box results (`whitebox.contrast.paired_contrast`). Mean +0.0233 $[+0.0017, +0.0493]$. Paired median +0.0025 $[+0.0000, +0.0094]$. Leave-two-out mean +0.0090 — the two largest rows carry 63% of the mean. Exact two-sided sign test $p = 0.020$ on 29 of 42 positive positive/negative non-tied rows (Wilcoxon $p = 0.006$). And the paired sign-flip permutation test on the necessity-vs-matched-control leg: $p = 0.081$ on $n = 37$. **The sign test and the magnitude-based permutation test disagree.** More rows move the right way than the wrong way, reliably; but the typical row moves by +0.0025 $[+0.0000, +0.0094]$, an order of magnitude below the mean, and deleting two of fifty rows leaves +0.0090.

For scale comparison, The citation's normalized intervention effect is 8% on the mean, **0.9% on the median and 3% leave-two-out**, each dividing like by like. It is a ratio of two restoration effects, not a share of internal evidence: the same-mask sufficiency of 0.8862 $[0.7762, 0.9618]$ against necessity 0.2179 $[0.1158, 0.3289]$ (§E.6) is what rules that reading out. The robust reading is *at most a few per cent on a typical image, with a mean pulled to 8% by two images*. Both mean and robust summaries are retained because they describe different features of the paired distribution.

Estimator	MagicBrush	Splice, hard paste	Splice, hard paste (n=150)	Splice, feathered
n	50	50	150	50
mean ΔNS cite	+0.0001 [−0.0250, +0.0263]	+0.0233 [+0.0017, +0.0493]	+0.0198 [+0.0096, +0.0316]	+0.0041 [−0.0162, +0.0235]
median ΔNS cite	−0.0027 [−0.0150, +0.0033]	+0.0025 [+0.0000, +0.0094]	+0.0010 [+0.0000, +0.0065]	+0.0000 [−0.0012, +0.0035]
leave-2-out mean	−0.0118	+0.0090	+0.0147	+0.0052
tail share of mean	13422%	63%	27%	−23%
sign test p (+/−)	0.480 (22/28)	0.020 (29/13)	0.003 (79/45)	0.551 (25/20)
Wilcoxon p	0.6051	0.0058	0.0001	0.4200
permutation p	0.9961	0.0589	0.0003	0.7061
mean ΔNS ceiling	+0.0649 [+0.0252, +0.1038]	+0.2856 [+0.2471, +0.3252]	+0.2813 [+0.2592, +0.3041]	+0.2629 [+0.2226, +0.3048]
median ΔNS ceiling	+0.0551	+0.2659	+0.2575	+0.2556
cite/ceiling, mean	0.1%	8.1%	7.0%	1.5%
cite/ceiling, median	−4.9%	0.9%	0.4%	0.0%
cite/ceiling, L2O	−17.3%	3.3%	5.3%	2.1%

Extended audit at $n = 150$. The table above prints the Wilcoxon and the paired permutation beside the four estimators of the previous paragraph — six in all, three of them p-values — and of those three the permutation test is the one the hard-paste citation leg does *not* clear at $n = 50$ ($p = 0.059$). (The median's bootstrap interval also includes zero at that n, and still does at the larger one; that is a statement about the typical image, not about significance, and it is reported

either way.) Since 150 eligible triples exist in the corpus we re-ran the identical audit on all of them — same checkpoint at the same revision, same `AuditConfig` (grid 4, mean occlusion fill, area budgets, control stride, fragmented and ceiling controls, both inpainters), same prompt version, greedy decoding, seed 0 — for 4419 local forward passes and 150 generations over 189 minutes of MPS wall clock with no network.

Cohort nesting and numerical agreement. The extension contains the smaller cohort and reproduces its stored outputs. `scripts/certify_audit_extension.py` checks both and writes `results/splice_audit_extension_certificate.json`. The 50 smaller-cohort `sample_id`s nest in the 150 — they are scattered through the draw rather than a prefix, because the smaller cohort's 50 rows were selected by the served run's eligibility rule — and on the 50 shared rows the two runs agree **bit-exactly**: 637 of 637 comparisons across 13 stored columns, maximum absolute difference 0, and the generated `explanation_raw` string identical on every row. That is agreement across two independent process launches on a different sample draw, which is a stronger reproducibility statement than re-running one command twice. The certificate certifies.

The result. All three p-values clear, the permutation test included: mean +0.0198 $[+0.0096, +0.0316]$, paired median +0.0010 $[+0.0000, +0.0065]$, leave-two-out +0.0147, tail share falling 63% $\rightarrow$ 27%, sign test $p = 0.003$ on 79 of 124 positive, Wilcoxon $p < 10^{-3}$, permutation $p < 10^{-3}$, and the stored sign-flip permutation leg $p = 0.001$ on $n = 123$. The direction and the mean survive at three times the n; **the median still does not exclude zero** (+0.0010 $[+0.0000, +0.0065]$), so the effect on a typical image remains unresolved. The corresponding normalized intervention effects are 7% on the mean, 0.4% on the median and 5% leave-two-out, against a detection AUC of 0.975 $[0.958, 0.989]$, a ceiling leg of 0.2813 $[0.2603, 0.3053]$ and a minimum detectable effect of 0.0159 (6% of the detection margin, against 12% at $n = 50$).

Residual bias and multiplicity. First, the estimator's bias on the 39 citations that miss the annotated edit entirely — rows where ΔNS ought to be exactly 0 — is -0.0139 $[-0.0279, -0.0026]$, an interval that *excludes* zero at this n where at $n = 50$ it did not (-0.0228 $[-0.0583, +0.0009]$). Necessity on those rows is exactly +0.0000 on every one of them, as the locality certificate requires, so the whole residual sits in the control leg (+0.0139); its sign runs against a spurious citation effect, so the citation number above is if anything conservative, but the residual is resolved from zero at the larger sample size and a sensitivity-corrected reading should subtract it. Second, this paper's standing claim that text overlap carries no causal information is at this n a claim that could have failed: 3 of the 36 tests in the declared family reach raw $p < 0.05$ (best `bertscore` at $p = 0.013$), where none did at $n = 50$. 0 survive BH over that family (minimum $q = 0.28$), so the claim stands under the correction we declared — but the raw evidence moved toward the association, and reporting only the corrected null would hide that.

Power limits of the MagicBrush audit. Section E.8 already bounds it, and the replication makes the bound concrete. On MagicBrush the smallest citation effect the design could detect was 0.0374, 80% of the model's entire detection margin, and the fraction of the ceiling effect the citation recovers came out at 0% $[-43\%, 55\%]$ — an interval so wide it excludes nothing. On splice the same design resolves 0.0346, 12% of a margin that is six times larger, and the same fraction is 8% $[1\%, 17\%]$. The corpus did not change the model; it changed the resolution of the instrument by roughly a factor of six, and the effect that appears is one MagicBrush could not have seen.

Restoration-effect magnitudes. The verdict leg is enormous here and is an identity rather than an estimate: restoring the true mask returns the authentic image bit-for-bit, so $\mathrm{Nec}(M^\star) = 0.2856$[0.2468, 0.3247] *equals* the whole detection margin 0.2856 $[0.2468, 0.3247]$ and its area-exact control moves the verdict by 0.0000 $[0.0000, 0.0000]$ — exactly zero, to the last reported digit, because paste-back over a control region on this corpus rewrites no pixel at all. The citation leg is 0.0233 $[0.0050, 0.0563]$. The paired within-image gap between them is 0.2623 $[0.2217, 0.3048]$, and the cited-region/full-region normalized intervention ratio is 8% $[1\%, 17\%]$. Against the oracle positive control run on these same images, where a faithful-by-construction detector scores 0.752 and a decoy-citing one 0.000, FakeShield sits at 0.09 of the way from floor to ceiling. The relevant comparison is quantitative: *on the hard-paste construction, the cited-region/full-region normalized intervention ratio is 0.9% on a typical image, 3% with the two largest images removed, and 8% in the mean.* This ratio is not an additive evidence allocation. Every number in this paragraph is

from the hard-paste corpus; the fourth column of the citation-robustness table above, and Table 30 in Section I, run the same instrument on the feathered construction of the same edits.

The chat model on the same corpus is what makes that fraction legible. A fraction is uninterpretable without knowing what any model manages here. `qwen2.5vl:7b`, grounded by *its own* emitted box on the identical 50 images, recovers 31% $[17\%, 44\%]$ of its ceiling ($\Delta\mathrm{NS} = 0.2216$[0.1262, 0.3359]) — roughly four times FakeShield's 8% $[1\%, 17\%]$, mean against mean. So 8% is not the corpus's ceiling on what a citation can achieve; it is this system's. On MagicBrush the same ratio for the same chat model is 17% $[-83\%, 109\%]$, an interval that spans impossible values in both directions, which is the corpus effect expressed as a statistic that refuses to be computed.

Detection and grounded-region comparisons are corpus-dependent. (i) Detection. On splice/copy-move, FakeShield reaches AUC 0.981 $[0.955, 0.999]$ versus 0.587 $[0.520, 0.663]$ on MagicBrush. Its AUC difference from `qwen2.5vl:7b` (0.992) is -0.011, an unresolved contrast rather than an equivalence result. On MagicBrush it trails by -0.281 $[-0.358, -0.198]$. The weaker detection result therefore applies to DALL-E-2 whole-canvas edits, not to every manipulation family. **(ii) The citation gap against the chat model is larger here, not smaller.** Paired on 50 images against `qwen2.5vl:7b`, FakeShield is worse on IoU (-0.227 $[-0.318, -0.137]$, $p < 10^{-3}$), worse on CF-Score (-0.203 $[-0.282, -0.129]$, $p < 10^{-3}$) *and* worse on the causal sub-score (-0.198 $[-0.303, -0.107]$, $p < 10^{-3}$) — where on MagicBrush the ΔNS contrast straddled zero. The grounder asymmetry of Section E.4 still runs against FakeShield and still bounds how much of the IoU and CF gaps to credit; it does not explain the ΔNS gap away, and on this corpus that gap is real.

One thing the corpus confirms outright. The outside-mask locality condition, which MagicBrush fails (Section E.13), holds exactly here. 13 citations miss the annotated edit entirely; on those rows necessity is 0.0000 $[0.0000, 0.0000]$ and 13 of 13 are the bit-exact identity. That is the hedged claim in the abstract, measured on both sides of its condition.

E.4 Is the citation-null falsifiable? A positive control for the grounder

The MagicBrush citation-null has a second explanation that no number in it can exclude: FakeShield's cited region is produced by *our* post-hoc grounder, so a grounder that cannot find the right region would return the same null no matter what the model wrote. A citation-null measured through an uncontrolled instrument is not evidence about the model. We therefore run the grounder against a known answer, through the *identical* code path the audit uses (`posthoc_region`, area budget 12%), on the same 50 images.

Grounded from	n	IoU vs M^*	IoU/ceil	recall	centroid hit
Oracle text (true class + true cell)	50	0.176 $[0.129, 0.238]$	0.425	0.383	0.40
direction only	50	0.176 $[0.129, 0.238]$	0.425	0.383	0.40
artifact class only	50	0.091 $[0.058, 0.144]$	0.182	0.155	0.20
Decoy text (farthest cell)	50	0.033 $[0.017, 0.066]$	0.074	0.067	0.06
Random window (same area)	50	0.097 $[0.067, 0.132]$	0.211	0.175	0.22
`fakeshield-v1-22b/DTE-FDM` (real explanations)	50	0.112 $[0.074, 0.170]$	0.242	0.197	0.26
`llava:13b` (real explanations)	50	0.091 $[0.058, 0.144]$	0.182	0.155	0.20
`llava:7b` (real explanations)	50	0.089 $[0.057, 0.142]$	0.176	0.149	0.20
`qwen2.5vl:7b` (real explanations)	50	0.089 $[0.056, 0.142]$	0.180	0.151	0.20

The arms are declared before the numbers. **Oracle** feeds the grounder a synthetic explanation naming the true artifact class and the true 3×3 cell of the edit, written in FakeShield's own prose register — the ceiling. **Decoy** is identical except that the named cell is the *farthest* one; if the grounder scored the same here, it would be ignoring the text entirely. **Random** is a same-area window placed uniformly — the floor. The two gates both pass: oracle beats random by +0.079 $[+0.026, +0.146]$ in IoU and beats the decoy by +0.143 $[+0.100, +0.206]$, both intervals excluding zero. The instrument is text-sensitive and does recover a named region, so the null is not automatic.

Substantial grounding error remains. Even handed the true cell in its own vocabulary, the grounder reaches only IoU 0.176 $[0.129, 0.238]$, which is 0.425 of the geometric ceiling a 12%-area region can attain against these masks, recovers 0.383 of the edited pixels, and puts its centroid

inside the true mask on 40% of images. Worse for any reading that credits the artifact vocabulary: naming the artifact class with *no* direction scores 0.091 $[0.058, 0.144]$; its contrast with the random window 0.097 $[0.067, 0.132]$ is unresolved, while naming the direction with no artifact class reproduces the full oracle score exactly. **The entire localisation ability of this grounder is its 3×3 directional prior; the artifact heatmaps contribute nothing measurable on this corpus.** That is a limitation of our instrument, stated plainly, and it caps what any citation-dependent number in this paper can mean.

A second, harsher control uses a ground truth the corpus does not supply: the served models emit their own boxes, so we can ask whether post-hoc grounding of a model's text recovers the region *that model itself* meant. It does not. Across the three served models the recovery is IoU 0.098–0.132 against a random-window floor of 0.119 $[0.099, 0.145]$ on the same pairs, and 0 of 3 of the three contrasts has an interval excluding zero. On chat-model prose the post-hoc grounder is at chance against the model's own intended region.

The real prose against the same floor, printed either way. Reading FakeShield's position *between* the oracle and the floor as evidence that "its prose carries genuine but partial directional information, more than the chat models' text" is not licensed by anything measured here. Every arm above grounds *oracle* text — text we wrote, which names the true cell — so it establishes that the instrument can find a region when the prose points at one, and nothing about whether a model's own prose points anywhere. The contrast that does bear on it is the real prose against the same-area random floor, paired on the same 50 images, and we print it here: FakeShield's own explanations beat that floor by +0.0146 $[-0.0337, +0.0796]$ IoU, an interval that includes zero, better on 18 images and worse on 19 (13 tied). **Through this grounder, the contrast between FakeShield's real prose and a random window of the same area is unresolved**, and the same holds for every chat model: 0 of 4 real-prose contrasts on this corpus have an interval excluding zero. What is left of the positive control is what it was designed to establish and no more — the citation-null is falsifiable as instrumented, because oracle text moves the grounded region (+0.079 $[+0.026, +0.146]$ over random, +0.143 $[+0.100, +0.206]$ over decoy) — and FakeShield's absolute position (0.112 $[0.074, 0.170]$, oracle minus FakeShield +0.065 $[+0.015, +0.126]$) must be read against that null result rather than as a weak version of the oracle's. This bounds every citation-dependent number in this paper from below and we state it in the limitation section as well.

Both gates and the whole loss profile reproduce on the splice corpus with the same instrument (`tables/table_grounder_control_splice.tex` in the artifact: oracle over decoy $+0.149$ $[+0.103, +0.206]$, oracle over random $+0.080$ $[+0.006, +0.154]$, with the artifact-class-only contrast against random again unresolved), so nothing in this subsection is specific to MagicBrush.

Consistent spatial parsing. The structured and free-text branches of `posthoc_region` both resolve location phrases through `detect_spatial_term`. This prevents an unrecognised noun phrase from silently reverting to a centre prior. All 200 stored audit rows use native regions or the free-text branch.

E.5 Swapping the instrument: a served VLM as the grounder

The subsection above leaves the citation leg resting on an instrument that is at chance on the input it is actually fed, and two hypotheses it cannot separate: **(a)** the models' citations carry no localisable signal, or **(b)** our grounder cannot convert a correct verbal citation into a region. We replace the instrument and hold everything else fixed.

What changed. Only the map from explanation text to region: instead of *free-text parser* $\rightarrow$ *artifact heatmap* $\times$ *directional prior* $\rightarrow$ *densest window*, we hand `qwen2.5vl:7b` the image and the explanation and ask for the box the explanation describes (`scripts/run_vlm_grounder_control.py`). The corpus, the `sample_id` set, the masks, the scoring function and *every arm* are the ones the post-hoc control used. Emitted boxes are rescaled about their centre to the same 12% area budget the post-hoc grounder is pinned to, because a served model answers with a box of whatever size it likes and IoU against a mask twenty times your area is bounded near zero whatever you point at.

The instrument is validated on MagicBrush — and it clears the gates by more than the post-hoc grounder does. Oracle prose beats a same-area random window by $+0.0938$ $[+0.0467, +0.1430]$ IoU (better on 34 images, worse on 10) and beats the decoy by $+0.1117$ $[+0.0596, +0.1598]$ (32 versus 5), against $+0.079$ $[+0.026, +0.146]$ and $+0.143$ $[+0.100, +0.206]$ for the post-hoc grounder. Oracle IoU is 0.183 against a random window's 0.089; 2 of 400 calls failed to parse.

And on real prose it still lands on the floor. Paired against the same-area random window on the same images: FakeShield $+0.0055$ $[-0.0448, +0.0582]$, `llava:13b` $+0.0324$ $[-0.0144, +0.0819]$, `llava:7b` $+0.0266$ $[-0.0153, +0.0690]$, `qwen2.5vl:7b` -0.0098 $[-0.0536, +0.0374]$. *0 of 4 real-prose contrasts have an interval excluding zero, under an instrument that demonstrably recovers the region when the prose names it.* On this corpus hypothesis (b) is refuted for a grounder strictly better than ours, and two instruments built on different principles agree: **neither resolves a citation effect on this corpus.** Because the two are built on different principles, the unresolved result is not attributable to the parser alone — but 0 of 4 real-prose contrasts exclude zero, and against a minimum detectable effect of 80% of the detection margin this is a power limit, not an established absence.

On the splice corpus the new instrument fails its own gates, and that is a result about the instrument rather than about any model. We ran it there too, and it does not validate: oracle prose does *not* beat a random window (-0.0205 $[-0.0770, +0.0380]$, better on 18 images and worse on 22) and does not beat the decoy ($+0.0074$ $[-0.0299, +0.0444]$). Its oracle IoU is 0.080 against a random window's 0.101 — pointed at the true cell in its own vocabulary, the served grounder does *worse* than chance placement. **No citation conclusion may be drawn on splice through this grounder**, and we draw none; the post-hoc grounder, whose gates that corpus does pass, remains the only validated citation instrument there. We report the failed control as an instrument limitation, not as evidence about the audited model's faithfulness.

What the two halves license, and a caution about instrument transfer. The gates are not a formality that a working instrument passes everywhere: the same grounder, same prompt, same area budget and same scoring passes decisively on one corpus and fails on the other. An auditing protocol that validates its instrument once and then reuses it across corpora would have reported the splice numbers as findings. Ours reports them as an instrument failure because the control was re-run per corpus, which is the practice we would ask of anyone adopting this protocol.

Two further notes. The post-hoc grounder's finding that its whole localisation ability is a directional prior does *not* carry over: under the VLM grounder the artifact class contributes materially (artifact-only 0.106, where-only 0.120, both together 0.183), so "the forensic vocabulary carries no location" was a statement about our parser. And this subsection does not make the grounded region *correct*: an oracle-fed IoU of 0.183 against a 12% budget is far from the geometric ceiling, so the power bound of Section E.8 stands. Agreement across instruments limits a parser-specific explanation, but does not remove shared grounding error or the power limitation.

E.6 Dissolving the grounder confound: the system's own localiser

Everything above measures FakeShield's citation through *our* post-hoc grounder, and the two subsections above establish that the grounder is at chance against a served model's own box. That is a confound of the worst kind, because it is not a wide interval but an unfalsifiable leg: a grounder that simply cannot find the right region returns the same null whatever the model wrote. The released repository contains the instrument that removes it. **MFLM** is FakeShield's own localisation module — a GLaMM/SAM grounded-segmentation model (`MBZUAI/GLaMM-GranD-Pretrained (GLaMM / SAM ViT-H)`) shipped in the same `zhipeixu/fakeshield-v1-22b` checkpoint. It is fetched, loaded under the same discipline as DTE-FDM, gated, and the citation leg is re-run through it on the identical certified-local corpus and the identical 50 `sample_id`s.

The load, gated the same way. MFLM's `config.json` names `GLaMMForCausalLM`, a class that exists only in a repository pinned to `transformers==4.28.0`. As with DTE-FDM we refuse to vendor it and remap instead, bijectively, onto *native* `transformers` containers: the language stack onto `LlavaForConditionalGeneration`, the grounding encoder onto `SamVisionEncoder`, `SamPromptEncoder` and `SamMaskDecoder`, and the text-to-SAM pro-

jection onto a plain `nn.Sequential`. 968 checkpoint tensors decompose as 893 mapped and 75 excluded, and the exclusions are enumerated with reasons rather than tolerated: 43 belong to GLaMM's bounding-box *input* branch, which the released `cli_demo.py` never reaches (it passes an empty box list), and 32 are `rotary_emb.inv_freq` buffers that `transformers` 4.56 recomputes. Of the 1284 live parameters, 893 come from MFLM and the remainder are the CLIP vision tower, which the checkpoint deliberately omits (`unfreeze_mm_vision_tower: false`; the released demo pulls it from the Hub) — and which we corroborate rather than assume, since the Hub tower is bitwise identical to the one embedded in the DTE-FDM checkpoint we already audit. 0 missing keys, 0 unmapped keys, 0 shape mismatches, 0 tensors left on `meta`; 27 of 27 fingerprinted tensors, spread across all four stacks, are **bitwise identical** to the checkpoint bytes at zero tolerance.

That is a bookkeeping proof, and bookkeeping does not establish that a rename table is *correct*. The decisive check is behavioural and adversarial to our own code: we load Meta's SAM implementation *as vendored in the released FakeShield repository* (`MFLM/model/SAM`, imported from a checkout and never copied into this project) and run it against our containers on the same weights and the same input. We state the provenance precisely because it bounds what the check proves: the reference path is independent of *our* rename table, which is the thing that could be wrong, but it is not an independent third-party reimplementation of SAM. The largest relative disagreement anywhere is 3.8e-04 (image encoder 3.9e-06, mask-decoder logits 2.8e-04), with 4 of 8 comparisons agreeing *exactly*; all agree within tolerance: yes. A wrong rename is an $O(1)$ effect, not a 10^{-4} one.

The acceptance test, run before the module was allowed into the audit. Table 8 is the gate. Against the true mask on the same 50 images, MFLM reaches IoU 0.4956 $[0.3928, 0.6007]$ (median 0.5197) where an area-matched random window reaches 0.1024 $[0.0657, 0.1550]$: paired, +0.3932 $[+0.2795, +0.5092]$, sign 38 of 50 positive, $p = 0.0003$. Two randomisation controls say the masks come from the released weights rather than from the architecture's shape prior — re-randomising the mask decoder costs +0.4224 $[+0.3165, +0.5301]$ and re-randomising the text projection +0.4388 $[+0.3345, +0.5442]$, both collapsing the module to the random floor. A purely geometric control that owes nothing to the fine-tuning (prompt SAM with a box and ask how much of the mask stays inside it) gives containment 0.6776 $[0.5601, 0.7777]$ against 0.2487 $[0.1977, 0.3221]$ through a randomised decoder, a paired +0.4335 $[+0.3190, +0.5463]$ ($p = 0.0002$). 6 of 6 reported criteria pass and the module is **accepted**. This is an *exploratory acceptance decision*, not confirmatory validation on an untouched set. **One of those criteria was rewritten after its first form failed, and we disclose it rather than report only the version that passed.** The box control initially demanded absolute containment ≥ 0.90; the measured value is 0.6776 $[0.5601, 0.7777]$, so that form failed. We replaced it with the comparison against the randomised decoder reported above, on the grounds that 0.90 was a threshold we invented with no basis in how SAM is trained — it is trained with jittered boxes and is not expected to respect them tightly — whereas the randomised-decoder contrast tests the property the criterion was for, namely that the geometry comes from the released weights. The absolute number is reported either way, and a reader who prefers the original threshold can see that it is not met. Two fidelity deviations are measured rather than argued: running the grounding stack in float32 instead of the released bfloat16 moves IoU by -0.0022 $[-0.0046, -0.0004]$, and the one prefix-space token `transformers` 4.56 inserts into a $\sim$309-token prompt moves it by -0.0076 $[-0.0157, -0.0006]$ (median +0.0000, $p = 1.0000$).

Reading Table 8. Rows 3–6 are the acceptance controls. *area* is the fraction of canvas each instrument spends, printed beside IoU because a like-for-like comparison must show what the IoU was bought with; our post-hoc grounder's area is fixed at 12% by construction, which is why its cell is not a measurement. The *shuffled narrative* row is the finding that runs against a natural reading of the others and is discussed below.

The native mask is not established as narration-conditioned. We ran one control that no acceptance criterion required, and it is the one that bounds what the rest of this subsection may claim. Feed MFLM image i together with a *different* image's narrative and it still produces essentially the same quality of mask: IoU 0.4198 $[0.3238, 0.5251]$ against the native 0.4956 $[0.3928, 0.6007]$, paired difference +0.0758 $[-0.0285, +0.1829]$, median +0.0005, sign 27 of 49 positive, $p = 0.5682$. **The shuffled-text control does not establish conditioning of MFLM's mask on the explanation DTE-FDM wrote.** It is an image-conditioned tamper segmenter that happens to be prompted with

Table 8: **THE CITATION INSTRUMENT, SWAPPED.** Every row is a way of turning FakeShield's output into a cited region, scored against the true mask on the identical 50 certified-local images.

Cited region from	n	IoU vs M^*	median	area
FakeShield's own MFLM mask	50	0.496 $[0.393, 0.601]$	0.520	0.121
our post-hoc grounder	50	0.126 $[0.091, 0.171]$	0.085	—
MFLM, another image's narrative	50	0.420 $[0.324, 0.525]$	0.312	0.118
area-matched random window	50	0.102 $[0.066, 0.155]$	0.018	0.122
MFLM, mask decoder randomised	50	0.073 $[0.064, 0.083]$	0.071	0.122
MFLM, text projection randomised	50	0.057 $[0.019, 0.140]$	0.000	0.117

prose. This matters for what the region *is*: it is the region the released system outputs and therefore the right region to intervene on, and it is emphatically not "the region the narration cites". Wherever this paper writes that the region leg is measured on the system's own region, that is the sense meant, and calling the MFLM mask a citation *of the explanation* would overclaim. The post-hoc grounder, for all its weakness, is the only instrument *in this FakeShield audit* that is a function of the narration at all — which is why it is reported beside the native mask rather than replaced by it. The second released system of Section E′ is the configuration where that instrument exists: its narration names anatomy, so the part names ground the region and an own-versus-shuffled contrast can be run at all. Whether the two systems would differ under a *common* region instrument is not something either audit measures.

(a) The two instruments, paired on the same images. The re-audit changes nothing but the map from output to region: same checkpoint, same `AuditConfig`, same corpus, same operator. That is checkable rather than assertable, and we checked it: the 5 legs that do not depend on the cited region ($p(x)$, $p(x_{\mathrm{src}})$, $p(A_{M^\star}(x))$, necessity at the true mask, and the true mask's area) are *exactly* equal between the two runs on 50 of 50 rows, maximum absolute difference 0.0. The runs are paired, not merely comparable.

Reading Table 9. The mean, the paired median, the leave-two-out mean and the exact sign test are printed together because they disagree here, and that disagreement is the result: the mean rises with an interval excluding zero while the median barely moves, which is the signature of a minority of rows carrying the effect. 3 of 8 legs survive BH, and none of them is a $\Delta\mathrm{NS}$ leg.

Table 9 gives every leg. Read honestly, in the order the numbers force. Localisation improves enormously and unambiguously: IoU +0.3693 $[+0.2608, +0.4780]$, median +0.3259, sign 41 of 50 positive, $q < 10^{-4}$, which rejects. Two distinct estimands must be separated. *Occlusion sensitivity at the citation* — the model's own suspicion of the region it cites, s_{cited} — falls by -0.0680 $[-0.1021, -0.0383]$ ($q = 0.0025$, rejects). *Specificity at the selected control* — a different estimand, on the 47 rows where an area-matched control exists — falls by -0.0112 $[-0.0218, -0.0038]$ ($q = 0.0020$, rejects). Both are consistent with the same geometry: a native mask is a tighter region, so both the cited region itself and an area-matched control drawn against it disturb less than the grounder's fixed window does. And the causal legs move in the direction the native mask predicts while failing every rank test: $\Delta\mathrm{NS}$ at the matched control is +0.0565 $[+0.0171, +0.0991]$ with an interval excluding zero, but its paired median is only +0.0076, its sign test is 28 of 47 positive ($p = 0.2430$), its leave-two-out mean drops to +0.0386, and it does **not** reject under BH ($q = 0.2777$). Necessity at the cited region behaves the same way (+0.0415 $[+0.0068, +0.0780]$, median +0.0042, sign 26 of 50 positive, $q = 0.8877$). **Both facts must be stated: the mean rises with an interval excluding zero, and no rank test survives.** As in the splice and white-box analyses, the shift is concentrated in a minority of rows.

(b) Localisation, with the canvas each instrument spent. The native mask reaches 0.4956 $[0.3928, 0.6007]$ against the grounder's 0.1263 $[0.0891, 0.1668]$, and the gain is not bought with area: MFLM spends 0.1213 $[0.0934, 0.1580]$ of the canvas against the grounder's fixed 12%, with true masks at 0.1560 $[0.1268, 0.1868]$. It also fails far less often: the cited region misses the true mask entirely on 1 row through MFLM against 13 through the grounder, of 50.

Table 9: **NATIVE MASK MINUS POST-HOC GROUNDER**, paired on the identical 50 images, Benjamini–Hochberg over a 8-member family frozen before this native-versus-grounder contrast was run.

Leg (native − grounder)	n	mean	median	L2O	sign	p	q
localisation IoU	50	+0.3693 $[+0.2608, +0.4780]$	+0.3259	+0.3436	41/50	$< 10^{-4}$	$< 10^{-4}$
occlusion sensitivity at the citation	50	−0.0680 $[-0.1021, -0.0383]$	−0.0325	−0.0523	13/50	0.0009	0.0025
necessity at the cited region	50	+0.0415 $[+0.0068, +0.0780]$	+0.0042	+0.0265	26/50	0.8877	0.8877
specificity at the selected control	47	−0.0112 $[-0.0218, -0.0038]$	+0.0000	−0.0055	3/47	0.0005	0.0020
ΔNS (matched control, any source)	47	+0.0565 $[+0.0171, +0.0991]$	+0.0076	+0.0386	28/47	0.2430	0.2777
ΔNS (congruent matched control)	25	+0.0790 $[+0.0139, +0.1498]$	+0.0210	+0.0458	17/25	0.1078	0.1903
ΔNS (fragment control)	47	+0.0509 $[+0.0142, +0.0914]$	+0.0101	+0.0345	28/47	0.2430	0.2777
cited area fraction	50	+0.0018 $[-0.0283, +0.0349]$	−0.0200	−0.0127	19/50	0.1189	0.1903

(c) The normalized intervention effect, recomputed through the system's own region — and it more than doubles. The quantity reported for FakeShield is the ratio between restoration at its output region and at the full edit. On the certified-local corpus the denominator is exactly computable, because paste-back over $M^\star$ returns the authentic image bit-for-bit and the denominator is therefore the whole detection margin rather than an estimate. We recompute it with the released code's ceiling-normalised implementation called *unchanged* — the same function and the same 0.05 denominator floor as the grounder leg, on the 49 of 50 rows the floor admits (1 dropped for denominator stability, by the same declared rule, from both arms). A separate, smaller exclusion applies to the controlled contrast above: 3 native masks (`splice_134290`, `splice_393224`, `splice_50151`, covering 34–58% of the canvas) admit neither a congruent nor a fragmented area-matched control, so 47 of 50 rows enter the ΔNS contrast; their uncontrolled necessities are in the rows file and none exceeds 0.051 in magnitude. The normalized effect is:

citation instrument	cited/full intervention ratio	median
our post-hoc text grounder	0.0920 $[0.0353, 0.1582]$	0.0025
FakeShield's own MFLM mask	**0.2179** $[0.1158, 0.3289]$	0.0344

paired +0.1259 $[+0.0083, +0.2471]$, median +0.0169 $[-0.0436, +0.1101]$, sign 26 of 49 positive ($p = 0.775$), Wilcoxon $p = 0.171$, permutation $p = 0.0515$, leave-two-out +0.0883.

Reading Table 10. Each row applies the same floor to both region instruments; intervals are over eligible paired images. The ratio is operational, not an additive evidence share.

Construction replication with the native output mask. The decisive construction check reruns MFLM itself, rather than transferring a post-hoc-grounder result across instruments. It requests the identical 50 sample IDs on the 3-pixel feathered corpus and retains 48 native masks, with no fallback; 46 rows enter the normalized effect at the unchanged 0.0500 floor. Table 11 reports every causal leg separately.

Reading Table 11. Every estimate carries the n it was computed on: necessity is defined on every audited row, the control-subtracted legs only where an area-exact control exists, and the normalized effect only above the denominator floor. Selected-control specificity is shown before subtraction, and the normalized ratio is not an evidence allocation.

The feathered raw necessity is +0.0417 $[+0.0126, +0.0733]$, its selected-control subtraction is +0.0457 $[+0.0168, +0.0776]$, and its normalized effect is 0.1221 $[0.0195, 0.2312]$ (median 0.0253). The denominator grid spans 0.01 through 0.1:

Every floor's eligibility and interval is printed rather than selecting the most favourable row.

Necessity and sufficiency on the same native masks. MFLM is deterministic, so we reconstruct every retained hard-paste mask from its frozen DTE-FDM narration and verify the rerun IoU against the audit (maximum discrepancy 0). Restoring the mask gives necessity; starting from the authentic source and inserting manipulated pixels in the same mask gives sufficiency. The latter is also subtracted against an equal-area disjoint translation.

Table 10: Denominator-floor sensitivity of the normalized intervention effect.

denominator floor	eligible n	MFLM output region	post-hoc grounder
0.010	50	0.227 $[0.126, 0.337]$	0.110 $[0.045, 0.186]$
0.025	49	0.218 $[0.116, 0.329]$	0.092 $[0.035, 0.158]$
0.050	49	0.218 $[0.116, 0.329]$	0.092 $[0.035, 0.158]$
0.075	47	0.223 $[0.118, 0.337]$	0.100 $[0.042, 0.170]$
0.100	44	0.233 $[0.124, 0.353]$	0.099 $[0.038, 0.174]$

Table 11: **NATIVE MFLM ACROSS CONSTRUCTIONS.**

construction	masks / requested	IoU	raw Nec	selected-control Spec	ΔNS	normalized effect
hard paste	50/50	0.4956 $[0.3919, 0.5999]_{n=50}$	+0.0720 $[+0.0379, +0.1098]_{n=50}$	−0.0036 $[-0.0070, -0.0010]_{n=47}$	+0.0804 $[+0.0452, +0.1202]_{n=47}$	0.2179 $[0.1158, 0.3289]_{n=49}$
3-px feather	48/50	0.4862 $[0.3773, 0.5951]_{n=48}$	+0.0417 $[+0.0126, +0.0733]_{n=48}$	−0.0023 $[-0.0058, +0.0010]_{n=46}$	+0.0457 $[+0.0168, +0.0776]_{n=46}$	0.1221 $[0.0195, 0.2312]_{n=46}$

Necessity is +0.0720 $[+0.0405, +0.1137]$ while sufficiency is +0.2514 $[+0.2089, +0.2924]$; after its own equal-area floor the sufficiency contrast is +0.2401 $[+0.1765, +0.3009]$. Their normalized effects are 0.2179 $[0.1217, 0.3361]$ and 0.8862 $[0.7762, 0.9618]$, on the same $n = 49$ denominator-eligible rows. This directly tests the redundancy and interaction alternative that a low necessity alone cannot distinguish.

How much of the low necessity is coverage. The normalized effect is a per-row ratio, so it can be read against each native mask's geometry. Coverage, the share of $M^\star$ inside the mask, is derived per row from the stored IoU and areas, $|R \cap M^\star|/|M^\star| = \mathrm{IoU}\,(|R| + |M^\star|)/\big((1 + \mathrm{IoU})\,|M^\star|\big)$; its median over the 49 denominator-eligible rows is 0.53. The ratio follows it: Spearman +0.76 $[+0.55, +0.88]$, with 0.45 $[0.29, 0.61]$ on the 25 rows at or above median coverage against -0.02 $[-0.07, 0.03]$ on the 24 below it (difference +0.47 $[+0.30, +0.64]$; Table 14 prints the 0.5, 0.8 and 0.9 cuts beside the median split). The feathered construction also shows the dependence (Spearman +0.70 $[+0.51, +0.80]$, $n = 46$), and normalized sufficiency on the hard-paste split is 1.03 $[1.00, 1.05]$ against 0.74 $[0.57, 0.90]$. On a hard-paste splice, restoring a mask that covers part of the edit leaves the rest of the edit and its seam in the image, so a low ratio on those rows is also what a detector that does use the edited pixels would return; the positive control's partial citations show necessity collapsing as a faithful oracle's box slides off the edit (Table 6). The overall mean mixes different geometric conditions: it mixes rows where restoring the output region moves the elicited verdict with rows where the measured change is near zero. The analysis was added after inspection (Table 5); the Spearman correlation is threshold-free and every cut is printed.

Against regions of the same coverage. Coverage leaves one question the stored rows cannot answer: is the native mask's effect what *any* region covering the same share of the edit would give, or does where the mask sits inside the edit matter? For each of the 49 rows, 3 connected sub-regions of $M^\star$ were grown by breadth-first search from seeded random pixels, each holding the native mask's coverage of $M^\star$, restored with paste-back and scored through the same elicited readout (`scripts/run_mflm_coverage_matched.py`, with analysis rules fixed before computation and post-conditions checking completion independently of the outcome). The rerun reproduced the stored $p(x)$ exactly on 49 of 49 rows. Outside $M^\star$ this corpus is certified bit-exact, so the mask's pixels outside the edit restore nothing and the contrast concerns the pixels inside it. The native mask's normalized necessity is 0.2179 against 0.2877 for the matched sub-regions, paired -0.0698 $[-0.1047, -0.0319]$ (permutation $p = 0.0004$, sign 35 of 48 negative), and -0.0891 $[-0.1306, -0.0499]$ on the 25 rows at or above median coverage against -0.0497 $[-0.1052, +0.0145]$ on the 24 rows below median coverage (Table 15). FakeShield's mask therefore carries less normalized necessity than connected regions of the edit with its own coverage: coverage alone does not account for how low its ratio is. Two limits stay attached. Each sub-region is grown from one seed while the mask's intersection with the edit need not be connected, so placement and shape are compared together; and the readout is the elicited one, not the released paragraph, which flips on 0 of 50 fully restored pairs.

Interpretation of the normalized effect. The upper end of the normalized-effect interval is below 0.5, but that does not prove the complement is a percentage of evidence elsewhere: redundancy and interactions make such an allocation invalid. The normalized effect rises from 9% through our

Table 12: **FEATHERED NATIVE-MASK DENOMINATOR SENSITIVITY.**

denominator floor	kept / paired n	normalized effect	median
0.010	48/48	0.1360 $[0.0292, 0.2453]$	0.0253
0.025	47/48	0.1389 $[0.0300, 0.2495]$	0.0327
0.050	46/48	0.1221 $[0.0195, 0.2312]$	0.0253
0.075	44/48	0.1122 $[0.0064, 0.2239]$	0.0103
0.100	44/48	0.1122 $[0.0064, 0.2239]$	0.0103

Table 13: **BIDIRECTIONAL INTERVENTIONS ON IDENTICAL NATIVE MASKS.** The two normalized rows use the same full-region denominator and floor; they are operational effects, not additive allocations.

direction on same native mask	n	mean [95% CI]	median
necessity (restore cited pixels)	50	+0.0720 $[+0.0405, +0.1137]$	+0.0120
sufficiency (insert cited pixels)	50	+0.2514 $[+0.2089, +0.2924]$	+0.2443
ΔSuf versus equal-area control	34	+0.2401 $[+0.1765, +0.3009]$	+0.2443
normalized necessity	49	0.2179 $[0.1217, 0.3361]$	0.0344
normalized sufficiency	49	0.8862 $[0.7762, 0.9618]$	1.0000

grounder to 22% through its own, demonstrating sensitivity to the region instrument. We report denominator-floor sensitivity and raw necessity, specificity and ΔNS separately. The paired difference's own robustness is the same story as (a): the mean rises with an interval excluding zero, the median moves by +0.0169 $[-0.0436, +0.1101]$ and no rank test clears 0.05.

What this settles and what it does not. It settles the falsifiability objection. The native-mask leg does not rely on the post-hoc grounder: it is measured through the localiser the audited system ships, which beats an area-matched random window by +0.3932 $[+0.2795, +0.5092]$ and collapses under both randomisation controls. A null measured this way is a null about the model. It does not settle the *narration* question, because of the shuffled-narrative result above: what is testable is the claim about the region the system outputs, not about the region its prose describes. And it does not rescue the composite or change any grounder-free number in this section, since those never involved a citation.

E.7 An adjudicated citation subset, and who adjudicated it

The citation subset provides an automated adjudication, not human validation. **The rater is the authoring system (a large language model), not an independent human.** The three labels and the decision rule were fixed before any row was read; the rater saw the editing instruction, the true 3×3 cell, the mask area and the model's explanation, and was blind to the grounded region and to every audit outcome; the complete stimulus set and every label with its rationale ship in `results/citation_adjudication_{sheet,labels}.json` so a human can re-rate the identical subset and diff the two.

On 20 FakeShield rows, the explanation's *text* names a location matching the true edit on 12 (60%), names a different one on 8 (40%), and names none on 0. Ambiguous rows were resolved in the model's favour, which is the conservative direction here: a higher match count strengthens the rival hypothesis that our grounder, not the model, is what fails.

Two things follow, and they point in opposite directions. First, **the grounder loses most of a correct citation**: on the 12 rows whose text a rater judged correct, the grounded region reaches IoU 0.048 $[0.012, 0.111]$ — below the 0.176 $[0.129, 0.238]$ the oracle text achieves and no higher than on the rows judged wrong (0.093 $[0.025, 0.230]$). The objection is substantiated: our citation instrument does not convert a correct verbal citation into a correct region. Second, and this is why the section's conclusion survives, **the causal null does not move when we condition on the text being right**: ΔNS on the match rows is -0.0016 $[-0.0514, 0.0479]$, still centred on zero. That is $n = 12$ with a correspondingly wide interval and we draw no strength from it beyond this: the null is not an artefact of having selected rows where the model said the wrong thing.

Table 14: **NORMALIZED NECESSITY BY COVERAGE OF $M^\star$.** Mean per-row ratio at FakeShield's native mask with 95% bootstrap intervals, split at each coverage cut.

construction	coverage split	n at/above	ratio at/above	n below	ratio below
hard paste	median (0.53)	25	0.45 $[0.29, 0.61]$	24	-0.02 $[-0.07, 0.03]$
hard paste	0.5	26	0.44 $[0.28, 0.59]$	23	-0.03 $[-0.08, 0.02]$
hard paste	0.8	20	0.57 $[0.42, 0.73]$	29	-0.02 $[-0.07, 0.02]$
hard paste	0.9	17	0.62 $[0.45, 0.78]$	32	0.00 $[-0.05, 0.06]$
3-px feather	median (0.71)	23	0.35 $[0.21, 0.49]$	23	-0.11 $[-0.19, -0.04]$
3-px feather	0.5	25	0.32 $[0.17, 0.46]$	21	-0.11 $[-0.20, -0.03]$
3-px feather	0.8	19	0.43 $[0.27, 0.58]$	27	-0.09 $[-0.17, -0.03]$
3-px feather	0.9	14	0.49 $[0.35, 0.64]$	32	-0.04 $[-0.13, 0.06]$

Table 15: **NATIVE MASK AGAINST SAME-COVERAGE SUB-REGIONS OF $M^\star$.** Mean normalized necessity per row; paired difference with 95% bootstrap interval; sign test.

rows	n	native mask	same-coverage sub-regions	native − matched	sign test
all	49	0.2179	0.2877	-0.0698 $[-0.1047, -0.0319]$	35 of 48 negative
coverage at/above median (0.53)	25	0.4493	0.5384	-0.0891 $[-0.1306, -0.0499]$	18 of 25 negative
coverage below median	24	-0.0232	0.0266	-0.0497 $[-0.1052, +0.0145]$	17 of 23 negative

Cross-model adjudication and rater disagreement. A single rater who is also the author of the hypothesis is the weakest possible adjudicator, so we added `gemma3:27b` and `mistral-small3.1:24b` — different model families, served locally, given the *identical* sheet and the identical decision rule including the instruction to resolve ambiguity in the model's favour (`scripts/run_independent_adjudication.py`, 0 unparseable replies of 20×2). They are automated raters of different provenance, not a human panel; the independence is architectural and we claim nothing more.

Text-location agreement. the two independent raters call 6 and 7 of 20 rows a location match where the authoring system called 12. They agree with *each other* almost perfectly (Cohen's $\kappa = +0.886$, raw agreement 0.95) and with the authoring system only fairly ($\kappa = +0.259$ and $+0.340$). Two raters that converge on each other and diverge from us, in the direction of us being more generous, is the signature of exactly the bias having the author rate their own subset invites. **The 60% text-match rate above should be read as an upper bound; independent raters put it near half that.**

Conditional intervention effects. ΔNS on the rows each rater independently called a match is -0.0016 $[-0.0520, +0.0479]$ (n=12), -0.0297 $[-0.0986, +0.0288]$ (n=6) and -0.0368 $[-0.0944, +0.0151]$ (n=7) — every interval contains zero (all 3 raters: yes). The sentence this subsection exists to support, *the causal null does not move when we condition on the text being right*, is therefore rater-invariant, although the matched subsets remain small and all raters are automated.

E.8 THE NULL, BOUNDED BY POWER RATHER THAN ASSERTED

An interval containing zero is not a proven absence, and it is easy to drift toward writing it as one. The scale that makes FakeShield's citation null readable is its observed full-restoration margin: whole detection margin is $p(x) - p(x_{\text{src}}) = 0.0466$[0.0103, 0.0867]. Against that, the citation ΔNS interval's upper end is 0.0275, i.e. 59% of that margin, and the smallest effect this design could reliably detect (paired, $n = 50$, two-sided $\alpha = 0.05$, 80% power) is 0.0374 — 80% of that margin. So the defensible statement is bounded, not absolute: *restoring the cited region moves this model's verdict by at most about 0.0275 in probability, and the experiment could not have detected anything smaller than roughly 80% of its observed full-restoration margin.* We do not claim the citation effect is zero. Its interval includes zero at this n, while the grounder-free ceiling effect 0.0649 $[0.0253, 0.1040]$ on the same rows, with the same n and the same instrument, *is* distinguishable. The contrast between those two — same images, same design, same power — is what the section rests on, and it is a contrast rather than an absence.

E.9 CONDITIONING ON DETECTION

Model	Stratum	n	IoU	Nec	dNS*	CF*	null
fakeshield-v1-22b/DTE-FDM	all	50	0.112	−0.030 [−0.054, −0.007]	0.000 [−0.024, 0.027]	0.056 [0.036, 0.098]	0.046
fakeshield-v1-22b/DTE-FDM	flagged	45	0.122	−0.031 [−0.057, −0.005]	−0.010 [−0.034, 0.017]	0.061 [0.039, 0.103]	0.046
fakeshield-v1-22b/DTE-FDM	detected	8	0.216	−0.065 [−0.158, 0.091]	−0.033 [−0.126, 0.107]	0.115 [0.039, 0.319]	0.046
fakeshield-v1-22b/DTE-FDM	separated	35	0.135	−0.035 [−0.068, −0.001]	−0.006 [−0.037, 0.029]	0.068 [0.041, 0.122]	0.046
llava:13b	all	50	0.105	−0.024 [−0.052, 0.005]	−0.004 [−0.025, 0.017]	0.052 [0.034, 0.079]	0.042
llava:13b	flagged	36	0.106	−0.013 [−0.038, 0.025]	−0.010 [−0.031, 0.007]	0.053 [0.032, 0.085]	0.042
llava:13b	detected	4	0.103	0.087 [−0.082, 0.319]	−0.039 [−0.147, 0.071]	0.046 [0.014, 0.094]	0.042
llava:13b	separated	33	0.106	−0.008 [−0.041, 0.032]	−0.004 [−0.028, 0.027]	0.053 [0.030, 0.092]	0.042
llava:7b	all	50	0.071	−0.003 [−0.010, 0.002]	−0.001 [−0.007, 0.003]	0.035 [0.022, 0.052]	0.026
llava:7b	flagged	50	0.071	−0.003 [−0.010, 0.002]	−0.001 [−0.007, 0.003]	0.035 [0.022, 0.052]	0.026
llava:7b	detected	0	–	–	–	–	0.026
llava:7b	separated	38	0.069	−0.003 [−0.011, 0.004]	−0.002 [−0.009, 0.003]	0.034 [0.020, 0.052]	0.026
qwen2.5vl:7b	all	50	0.206	0.029 [−0.023, 0.087]	0.016 [−0.057, 0.092]	0.106 [0.075, 0.146]	0.052
qwen2.5vl:7b	flagged	40	0.193	0.044 [−0.011, 0.116]	0.016 [−0.063, 0.103]	0.109 [0.075, 0.158]	0.052
qwen2.5vl:7b	detected	30	0.227	0.060 [−0.012, 0.149]	0.018 [−0.086, 0.131]	0.129 [0.086, 0.183]	0.052
qwen2.5vl:7b	separated	46	0.221	0.033 [−0.023, 0.094]	0.017 [−0.061, 0.101]	0.113 [0.081, 0.156]	0.052

The pre-registered `detected` stratum ($p(x) > 0.5$ and $p(x_{\text{src}}) \leq 0.5$) retains only 8 of 50 FakeShield rows, because the model's operating point is not centred on 0.5 — it flags almost everything. On those rows CF-Score is 0.115 $[0.039, 0.319]$: higher, and far too wide to carry a claim. We therefore also report a threshold-free stratum, `separated` ($p(x) > p(x_{\text{src}})$), which uses the pairing the corpus already provides and retains 35 rows; there CF-Score is 0.068 $[0.041, 0.122]$ and ΔNS is -0.006 $[-0.037, 0.029]$. Conditioning on detection moves the point estimate up and the conclusion not at all: on no stratum does ΔNS separate from zero.

E.10 AGAINST THE GENERAL-PURPOSE MODELS IT IS SUPPOSED TO BEAT

FakeShield −	llava:13b	llava:7b	qwen2.5vl:7b
n	50	50	50
d IoU	+0.007 [−0.043, +0.054]	+0.041 [−0.006, +0.096]	−0.094 [−0.169, −0.025]
d Nec	−0.006 [−0.023, +0.015]	−0.026 [−0.050, −0.006]	−0.059 [−0.114, −0.008]
d dNS*	+0.004 [−0.024, +0.038]	+0.001 [−0.023, +0.029]	−0.016 [−0.092, +0.067]
d dNS ceil	+0.0302 [−0.0091, +0.0648]	+0.0524 [+0.0126, +0.0906]	−0.0331 [−0.0910, +0.0305]
d CF*	+0.004 [−0.022, +0.037]	+0.021 [−0.002, +0.056]	−0.050 [−0.086, −0.008]
p(CF*)	0.808	0.155	0.014
d AUC	+0.020 [−0.049, +0.089]	−0.040 [−0.128, +0.046]	−0.281 [−0.358, −0.198]

Paired on 50 images against `qwen2.5vl:7b`, FakeShield is worse on localisation (-0.094 $[-0.169, -0.025]$, $p = 0.012$), worse on CF-Score (-0.050 $[-0.086, -0.008]$, $p = 0.014$), while its causal-sub-score contrast is unresolved (-0.016 $[-0.092, 0.067]$, $p = 0.700$). Against the two LLaVA models every citation-dependent contrast is unresolved, not equivalent. The row that deserves naming is `llava:13b`, the closest thing here to a scale-matched untrained baseline — same parameter scale, same base family, no forensic fine-tuning at all. Every citation-dependent contrast against it has an interval straddling zero: IoU $+0.007$ $[-0.043, +0.054]$, CF-Score $+0.004$ $[-0.022, +0.037]$ ($p = 0.808$), ΔNS $+0.004$ $[-0.024, +0.038]$. We do *not* read this as "forensic fine-tuning buys nothing": these are not the same base checkpoint, the grounder asymmetry runs against FakeShield, and $n = 50$ intervals this wide would not detect a moderate effect. We read it as the absence of evidence that the fine-tuning bought faithfulness of citation, on the one axis this paper can measure. The honest reading, given the grounding asymmetry, is not that FakeShield explains *less* faithfully than a chat model — the IoU and CF gaps are exactly where a post-hoc grounder would cost it — but that it does not explain *more* faithfully, and that ΔNS, the quantity the grounder confound bears on least, is null for all four.

E.11 TEXT-METRIC ASSOCIATIONS ON THE RELEASED FORENSIC SYSTEM

Text metric	fakeshield-v1-22b/DTE-FDM	llava:13b	llava:7b	qwen2.5vl:7b	degenerate (FS)
bertscore	+0.02 [−0.26,+0.30]	+0.04 [−0.24,+0.31]	+0.04 [−0.24,+0.32]	+0.05 [−0.23,+0.32]	False
bleu4	−0.09 [−0.36,+0.19]	+0.13 [−0.15,+0.40]	+0.18 [−0.11,+0.44]	+0.03 [−0.25,+0.30]	False
cider	+0.02 [−0.26,+0.30]	+0.16 [−0.12,+0.42]	+0.17 [−0.11,+0.43]	+0.37 [+0.09,+0.59]	True
cider_raw	+0.02 [−0.26,+0.30]	+0.16 [−0.12,+0.42]	+0.17 [−0.11,+0.43]	+0.37 [+0.09,+0.59]	True
llm_judge	−0.26 [−0.51,+0.02]	+0.35 [+0.07,+0.58]	−0.00 [−0.28,+0.28]	−0.09 [−0.36,+0.19]	False
llm_judge_served	+0.24 [−0.05,+0.49]	+0.21 [−0.08,+0.46]	+0.21 [−0.08,+0.46]	+0.48 [+0.22,+0.68]	False
rouge_l	+0.06 [−0.22,+0.34]	+0.10 [−0.19,+0.37]	+0.13 [−0.15,+0.40]	−0.04 [−0.32,+0.24]	False

Reading Figure 8. Each panel prints ρ with its interval, and, because CF-Score is a near-monotone function of IoU on these rows (Appendix O), the IoU-only correlation beside it. The dashed line is that model's own area-matched random-citation null. The point of the top band is that the field's standard instruments do not order these explanations by whether the cited artifact drives the verdict — and the `llm_judge` panel is the sharpest case, where the association with CF-Score is if anything *negative*.

This is the paper's headline correlation claim evaluated on a member of the population it indicts. Across 7 text metrics (2 of which are degenerate on this model and are reported as such), the strongest association with CF-Score is `llm_judge` at -0.26 $[-0.51, +0.02]$, $p = 0.070$; no non-degenerate metric reaches $p < 0.070$.

Degeneracy limits the served-judge comparison. That instrument is `llm_judge_served`, a served `qwen2.5:7b` rubric grader, and the degeneracy screen we apply to every other text metric flags it as **degenerate on 3 of 3 served models** — dynamic range 3.6e-05, i.e. the metric never fires there. Counted as events rather than as means: the judge answered YES on 0 of 150 served explanations and 4 of 50 of FakeShield's. The "0.080 $[0.020, 0.180]$ against 0.00000" gap is four YES answers on one side against a floor of floating-point dust on the other, and does not support a graded ranking.

What the screened instruments actually say is less convenient and more interesting. Paired per image across all 50 rows, on the *deterministic* rubric judge — which is non-degenerate on all four models and is therefore the text-side instrument this paper can defend — FakeShield scores 0.443 $[0.428, 0.474]$ against `llava:7b`'s 0.719 $[0.674, 0.771]$, a paired difference of -0.276 $[-0.329, -0.227]$. It is *below* all 3 served models, with all 3 intervals excluding zero; the smallest gap is -0.139 $[-0.156, -0.110]$. BLEU-4 and ROUGE-L agree with the rubric judge. BERTScore does the opposite: it puts FakeShield first on all 3 pairs (+0.033 to +0.062, intervals excluding zero). So of the 4 usable text metrics, 3 (`bleu4`, `llm_judge`, `rouge_l`) rank the released forensic model *last* and 1 (`bertscore`) ranks it *first*, on the same explanations against the same references. The field's text-side instruments do not agree with each other about which system reads best, none of them is measuring whether the cited artifact drives the verdict — and the served judge is too degenerate to support that comparison.

E.12 COMPOSITE-SCORE REDUNDANCY

The composite adds little to localisation on this corpus. On FakeShield's rows $\rho(\mathrm{CF}, \mathrm{IoU}) = +0.996$ — the composite is a rescaled IoU here too, for the same reason as everywhere else: $\Delta\mathrm{NS}$ sits at zero, so the sigmoid is close to 0.5 and only the localisation factor varies. The pair we ship (necessity under paste-back, beside IoU) is what carries this section; the composite is reported only so the reader can see it add nothing.

E.13 OUTSIDE-MASK RESIDUALS ON MAGICBRUSH

This section's instrument gave us a way to test the paste-back identity claim directly, and the test does not come out clean on this corpus. 22 of the 50 FakeShield rows cite a region with *zero* overlap with the annotated edit. On these rows, restoring the cited region would have zero effect if the annotated mask captured every pixel difference. It is not: it is -0.0067 $[-0.0142, -0.0024]$, an interval excluding zero. The cause is not the operator but the corpus, and the direction of the asymmetry is worth stating precisely. The stored *authentic* image is the original MS-COCO photograph; the stored *manipulated* image is a DALL-E-2 inpainting output that re-renders the whole canvas at its

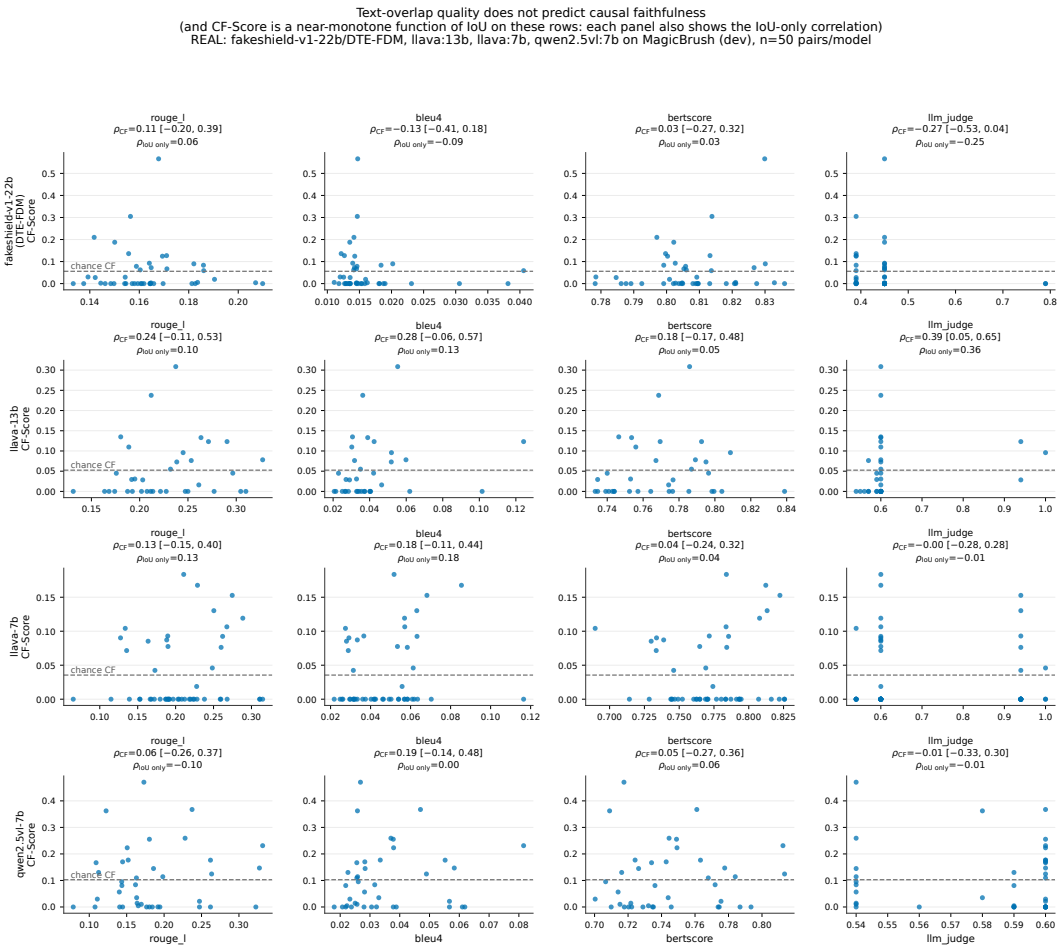


Figure 8: Text-overlap quality against causal faithfulness, per row, with FakeShield v1-22B (DTE-FDM) in the top band and the three served chat models below it on the identical pairs.

own native resolution (Section K). The manipulated image is therefore not "the photograph with a patch replaced", so pasting the photograph's pixels back over any region also removes that render's noise from it — on these rows it changes 0.100 $[0.093, 0.104]$ of the pixels in the region by a mean of 1.84 $[1.54, 2.30]$ intensity levels, and 0 of the 22 restorations leave the input unchanged. Three things bound it, and we state all three rather than the convenient one. The bias is small and *negative*, so on the raw necessity leg it works against finding a citation effect rather than manufacturing one. It applies to the control leg too, which is why every headline quantity in this section is ΔNS rather than raw necessity — but *how much* of it cancels is a measurement, not a lemma, reported below. And Section I removes it entirely on a second corpus, where the two images *are* bit-identical outside the mask, by certificate rather than by measurement. What is not defensible is the unqualified sentence "paste-back is the identity". Authentic-input identity holds on both corpora. Identity outside the *annotated mask* additionally requires bit-exact source/target agreement there; that condition fails on MagicBrush because of its whole-canvas re-render. The abstract and Section B carry the hedged form. Section K carries the same caveat for the served arm, on the same corpus.

Residual after control subtraction. The rows above are the natural test bed for the cancellation claim itself, because on them the citation carries no true evidence by construction, so ΔNS ought to be exactly 0 whatever the canvas-wide residual does. Whatever it is instead *is* the estimator's bias, and its sign says which way the instrument errs. On the 22 zero-overlap MagicBrush rows it is $\Delta\text{NS} = +0.0103$ $[-0.0131, +0.0355]$, decomposed as necessity -0.0067 against area-matched specificity -0.0164: the two legs are the same sign and the same order of magnitude, so they partly cancel — but *not* mostly. What survives is larger in magnitude than the necessity leg it is built

from (-0.0067) and of the same order as the specificity leg (-0.0164), and **it is signed *toward* a spurious citation effect**, not against one. Three things follow and none of them is comfortable. The interval includes zero, so this is a bound rather than a demonstrated bias. Its magnitude, $+0.0103$, is about 44% of the mean splice citation effect (+0.0233 $[+0.0017, +0.0493]$) and roughly four times its median (+0.0025 $[+0.0000, +0.0094]$) — which is one more reason the MagicBrush citation null (+0.0001 $[-0.0250, +0.0263]$) should never have been read as an absence, and one more reason the citation result is quoted from the corpus where the certificate makes this residual identically zero rather than from the one where it has to be estimated. On the hard-paste splice corpus the same measurement gives $\Delta\mathrm{NS} = -0.0228$ $[-0.0583, +0.0009]$ on 13 rows, signed *against* a spurious effect: there necessity on those rows is +0.0000 — exactly zero, as the certificate requires, since paste-back over a region disjoint from $M^\star$ rewrites no pixel — while the area-matched control leg is +0.0269, because a control chosen to be disjoint from the *cited* region may still intersect the true mask. That is a conservative bias, and it is the one that applies to the number the abstract quotes.

E.14 Native prose versus forced-choice readout

FakeShield exposes a paragraph, not a continuous native verdict probability. The primary audit therefore appends a one-word instruction to the released question and computes the exact full-vocabulary first-token softmax. We test that instrument on the same $n = 50$ hard-paste pairs before and after full restoration. Target prose and yes/no scores are reused byte-for-byte from the primary audit; only the restored prose and two alternative verbalizers require new calls. An explicit verdict is parseable on 50 targets and 50 restored images.

Native decisions and transitions. The number of restored images called manipulated is distinct from the number whose verdict changes from manipulated to authentic. Table 16 counts both directly from the stored per-image labels. The prose calls the fully restored image manipulated on 50 of 50 pairs (and the target on 50 of 50); it changes from manipulated to authentic on 0.

The table does not turn a prompted probability into a latent internal score. It quantifies when each verbalizer gives the same binary decision as the prose, and the paper's causal contrasts remain claims about the elicited readout. Its "native positive" column is constant down the table — 1.000 on both image states, in every verbalizer's row — because the native verdict is parsed from the *same* paragraph regardless of which verbalizer the elicited column uses; only the "forced positive" column varies. That constant column and Table 16's top-left cell are the same fact counted two ways.

The headline effects, recomputed under each verbalizer. Agreement with the prose is not the quantity the paper claims. So the normalized necessity and sufficiency are recomputed *themselves* under the two alternative verbalizers, on the same native MFLM masks, the same 49 denominator-eligible rows, the same paste-back operator and the same 0.05 ceiling floor; MFLM is rerun deterministically from each frozen narration and the rerun mask is checked against the audit's stored IoU on every row. The primary yes/no values are *reused*, not recomputed, ensuring a common primary baseline. Necessity is 0.2179 $[0.1217, 0.3361]$ on the primary readout, 0.1942 $[0.0936, 0.3128]$ under `manipulated`/`authentic` and 0.2654 $[0.1619, 0.3884]$ under `fake`/`real`; sufficiency on the same masks is 0.8862, 0.8731 and 0.8470. Sufficiency exceeds necessity on every readout: yes. Necessity's interval excludes zero on every readout: yes. Table 18 prints the ceilings, which differ by verbalizer and are why the raw effects are not comparable across rows of that table while the normalized ones are.

What this does and does not settle. The three verbalizers disagree sharply about the *decision*: on restored images their forced-positive rates are 0.840, 0.620 and 0.180. They nevertheless agree about the *effect*, which is the quantity this paper claims. So the readout is not interchangeable with the native paragraph — that remains true, and 0 of 50 restored pairs flip in prose — but the normalized necessity and sufficiency reported here are not artefacts of the one verbalizer chosen. Their ceilings differ because a verbalizer with a different operating point separates target from source by a different margin; normalizing by that margin is what makes the three rows comparable, and is why the raw effects in the same table are not. One bookkeeping note: the primary row's point estimate equals the primary audit estimate by construction, because its per-row values are reused rather than recomputed, but the interval printed here (0.1217–0.3361) comes from this table's own

Table 16: **NATIVE-PROSE VERDICT TRANSITIONS**, counted from the stored per-image verdicts, not from a rate. Empty cells print as zeros.

native verdict on target	native verdict after full restoration	pairs
manipulated	manipulated	50
manipulated	authentic	0
authentic	manipulated	0
authentic	authentic	0
total pairs		50

Table 17: **READOUT VALIDATION ON THE PRIMARY CORPUS.** Agreement is with the explicit verdict parsed from FakeShield's native paragraph; intervals are Wilson 95% intervals.

verbalizer	image state	parseable n	agreement	native positive	forced positive
yesno	target	50	1.000 $[0.929, 1.000]$	1.000	1.000
yesno	restored	50	0.840 $[0.715, 0.917]$	1.000	0.840
manipulated_authentic	target	50	1.000 $[0.929, 1.000]$	1.000	1.000
manipulated_authentic	restored	50	0.620 $[0.482, 0.741]$	1.000	0.620
fake_real	target	50	0.980 $[0.895, 0.996]$	1.000	0.980
fake_real	restored	50	0.180 $[0.098, 0.308]$	1.000	0.180

bootstrap over those rows so that all three readouts pass through one estimator; §3.3 quotes 0.1158–0.3289 from `analyze_mflm_citation.py`'s bootstrap over the identical rows.

E.15 Threats specific to this section

One checkpoint. FakeShield is one released forensic VLM; ForgeryGPT and SIDA are not audited and nothing here transfers to them by assumption. SIDA's weights are public (`saberzl/SIDA-7B`), so its audit is compute, not access (Section K). FFAA is audited separately (Section E′) under its released prompt, held verbatim from its repository, and on its own corpus, region instrument and readout the result runs the other way — a second audited configuration, not a verdict that FFAA is the more faithful model. **The narration and the mask are different objects.** MFLM, FakeShield's localisation module, is loaded, gated and audited (Section E.6), so the native-mask leg does not rely on our grounder — but the shuffled-text control does not establish that MFLM's mask is conditioned on what DTE-FDM wrote (+0.0758 $[-0.0285, +0.1829]$ against another image's narrative), so the re-audit is evidence about the region the *system* outputs and not about the region its *prose* describes. **The two classes differ by more than the edit.** On MagicBrush the authentic side is a photograph and the manipulated side is a whole-canvas DALL-E-2 render (native sizes differ on 49 of 50 audited pairs; outside-mask MAD median $1.53/255$, never 0), so a detector prompted about manipulation can earn AUC from render statistics alone and "the verdict is anchored to the edit" should be read as "part of it is" (Section K). This is the single largest reason the splice replication of Section E.3 matters. **The verdict readout is elicited.** Section E.14 measures agreement with the native paragraph and two alternative token families rather than assuming interchangeability. $n = 50$**.** The MagicBrush estimates use 50 paired images, with wide intervals: the grounder-free $\Delta\mathrm{NS}$ interval excludes zero, while the citation interval contains it. The latter is an unresolved effect, not a demonstrated absence. Other cohorts and estimators are identified beside their results.

E′ The second released system: FFAA on a face corpus that holds the counterfactual

This section audits the released FFAA checkpoint (Huang et al., 2024) with its verbatim user question, a generic LLaVA system wrapper, and a teacher-forced answer-field readout. The historical wrapper differs from FFAA's released forgery-analysis system prompt; these observations therefore measure this specified configuration, not the complete native inference pipeline. The

Table 18: **THE HEADLINE EFFECTS UNDER EACH ELICITED READOUT**, same masks, same rows, same operator. † reused from the primary audit.

elicited readout	n	ceiling	normalized necessity	normalized sufficiency
yesno†	49	+0.2856 [+0.2468, +0.3247]	+0.2179 [+0.1217, +0.3361]	+0.8862 [+0.7762, +0.9618]
manipulated_authentic	49	+0.4459 [+0.3910, +0.4981]	+0.1942 [+0.0936, +0.3128]	+0.8731 [+0.7742, +0.9381]
fake_real	49	+0.4985 [+0.4414, +0.5464]	+0.2654 [+0.1619, +0.3884]	+0.8470 [+0.7406, +0.9205]

three-candidate MIDS selector was also absent. Every number below is a generated macro from `results/real_ffaa_*.json`.

The preconditions, all four settled from files. Weights: `thu-huangzc/ffaa-mistral-7b` is held (15 GB, Apache-2.0). Loader: the checkpoint names the same `llava_llama` stack as DTE-FDM on a Mistral-7B language model, so the rename of Section E applies unchanged (686/686 tensors, 0 missing, 0 unmapped, 0 shape mismatches) and the bitwise gate runs on every load (10/10 fingerprints bit-identical, 686/686 tensors, gate passed). Corpus: the region-replacement face corpus of Appendix K (150 rows, outside-mask difference 0 on every row), so paste-back over $M^\star$ returns the authentic source. Prompt: the released repository ships ten paraphrases of one question (`playground/prompts.txt`, commit `a2be084`) and its inference code draws one uniformly per image; an audit must hold the prompt fixed across a sample's counterfactuals, so we fix paraphrase 0, "The image is a human face image. Is it real or fake? Why?", and measure the other nine below. The released code decodes greedily at temperature 0 and parses the answer into `Image description` / `Forgery reasoning` / `Analysis result` / `Probability` / `Forgery type`; both are reproduced from the fetched files, and a unit test compares the ten prompts against the fetched file byte for byte. Matching the user-question file does not establish conversation-template parity: the historical system line was generic LLaVA, whereas the released `conv_templates['v1']` uses a specialized face-forgery-analysis system line.

A readout the released format forces, and its validation. FakeShield answers a one-word forced-choice suffix at its first token; FFAA does not. Under the released question plus any one-word suffix, FFAA's first generated token is `Image` with probability above 0.99: the model is locked to its released answer format, and the `fake`/`real` families hold under 1% of the mass, so a renormalised ratio would be a ratio of residuals. We therefore teacher-force the assistant turn to begin at the released format's own verdict field (`Analysis result:`) and read the exact softmax at the next token over `_fake` against `_real`, `_Real`, every one a single token on the Mistral vocabulary; the description and reasoning fields are skipped, so the verdict is read from the pixels and the released question alone. The mass outside the two families is 0.0023 on average, and the graded decision agrees with the discrete `Analysis result` the model writes in its own greedy answer on 100.0% of 150 audited images (all 150 called fake natively; 100.0% of answers carry every released field). Not run, and stated rather than smoothed over: the released pipeline's two hypothetical follow-up prompts and its MIDS answer selector (which also returns a selected-answer forgery score), and the optional `dlib` face crop — the corpus images are aligned CelebA-HQ face crops, which is the released `-crop 0` path, and holding the crop fixed is what keeps the intervention to the pasted pixels alone.

Detection first. FFAA is a strong in-domain detector on this construction: paired AUC 0.997 [0.993, 1.000], flagging 100.0% of manipulated and 28.0% of authentic images, with $p(\text{fake})$ at a median of 1.000 on manipulated images (98% at or above 0.99) and 0.040 on their sources. A hard-paste seam is easier than a modern swap's, so this is not evidence about wild face-swap detection; the faithfulness question conditions on detection and is unaffected.

The verdict is anchored to the swapped pixels, with no grounder in the loop. Restoring $M^\star$ moves the verdict by $\mathrm{Nec}(M^\star)$ +0.709 [+0.648, +0.765] (median +0.955), and on this corpus the control leg is exact: paste-back over an area-matched control elsewhere rewrites no pixel, so its specificity is +0.0000 on 150 rows (largest magnitude 0.0000) and the ceiling $\Delta\mathrm{NS}$ is +0.709. The estimator's identity holds to the bit: restoring $M^\star$ reproduces $p(x_{\mathrm{src}})$ with mean absolute residual 0.0000.

Through our post-hoc grounder the narration is worse than a random window — and here that is a statement about the instrument. FFAA's `Forgery reasoning` names face parts; our post-hoc grounder maps artifact and spatial words to a compact 12% window. On this corpus the swapped region is small and non-rectangular (mean 0.120 of the canvas for the window against a part mask a third of that size), and the grounded window reaches IoU 0.047 [0.034, 0.062] (median 0.000) against $M^\star$. The necessity at that window is +0.013 [+0.004, +0.042] (median +0.0000), ΔNS is +0.007 [$-$0.010, +0.028] (median +0.0000), and the window/full-region normalized intervention ratio is 0.016 [0.005, 0.047] on the mean and 0.000 on the median ($n = 137$ rows with a stable denominator); this is not an additive evidence allocation. CF-Score is 0.025 [0.018, 0.033] against this model's own random-citation null of 0.023. On the oracle positive control run on the same corpus (10 arms), even the faithful oracle's own box reaches IoU 0.254 while its necessity is 0.991, so the grounder leg cannot separate "the narration does not cite the evidence" from "a fixed-window grounder cannot place a face part" here, as Section E.5 found on the splice corpus. We therefore do *not* read this null as a finding about FFAA's narration, and build the instrument the narration calls for.

The instrument a face narration calls for: the parts it names, on a corpus where naming them is not free. FFAA's `Forgery reasoning` names face parts. On the eight-part corpus every row swaps the same parts, so "the eyes and mouth look unnatural" is right on every image and cannot be scored. We therefore built a single-part corpus (150 rows; one of eyes, nose or mouth swapped per row, 50/50/50; same construction, same certificate, per-part masks stored) and replaced the grounder by the narration itself: the cited region is the union of the target image's masks for the part groups the narration names (eyes and brows; nose; mouth, lips and teeth). No learned component sits between the narration and the region. 145 of 150 narrations name at least one part (5 fall back to the post-hoc grounder and are flagged). Detection stays strong but is *not* unchanged: AUC 0.923 [0.895, 0.949] here against 0.997 on the eight-part corpus, an interval that excludes the eight-part estimate — one swapped group is a smaller edit than eight. The graded readout agrees with the written verdict on 94.0% of images.

The named region carries the verdict. Against $M^\star$ the named region reaches IoU 0.672 [0.620, 0.721] (median 0.619). Restoring it moves the verdict by +0.635 [+0.573, +0.693] (median +0.801) against a ceiling of +0.641 [+0.579, +0.698] (median +0.804): the named-parts/full-region normalized intervention ratio is 0.992 [0.961, 1.000] on the mean and 1.000 on the median ($n = 133$), ΔNS is +0.635 [+0.573, +0.693] (median +0.801), and CF-Score is 0.591 [0.540, 0.639] against a random-citation null of 0.336. On the oracle positive control run on the same corpus, the faithful oracle's necessity is 0.991 and the correct-but-non-causal arm's is 0.000 at identical IoU (0.509 against 0.509), so the instrument separates the two on this corpus as it did on the others.

And the naming is conditioned on the image: the shuffled-narration control. A narration that always says "eyes" would score perfectly on the eyes rows for free. We score each row with the narration written for a *different* row (a seeded derangement of the 150 narrations). FFAA names the swapped group on 96.0% of images against 54.0% under the shuffled narration (difference 42.0%, sign 67 of 71 positive, $p < 10^{-4}$); both name 1.69 groups per image on average, as a permutation of the same narrations must, so the gap is not verbosity. Per swapped group: eyes named on 100% of eyes rows against 42% of the others; nose 98% against 16%; mouth 90% against 51%. The causal half is paired on the same images: restoring the parts the image's own narration names moves the verdict by +0.652 (median +0.855) against +0.352 (median +0.035) for the parts a shuffled narration names — paired +0.300 [+0.234, +0.369], median +0.000 (the paired difference is exactly 0 wherever the shuffled narration happens to name the swapped group, which it does on 54.0% of rows, so the median is not the statistic to read here), sign 61 of 62 positive, Wilcoxon $p < 10^{-4}$, permutation $p < 10^{-4}$, leave-two-out +0.290, against a ceiling of +0.657 ($n = 140$ of 150; the 10 rows without both legs are *all* mouth rows — 5 because FFAA's own narration named no part and 5 because the shuffled one did not — so the pair is missing exactly where naming is weakest, which is stated here rather than left inside an n); by group, eyes +0.593 vs +0.359 (paired +0.234 [+0.134, +0.342], $p < 10^{-4}$), nose +0.739 vs +0.359 (+0.380 [+0.262, +0.500], $p < 10^{-4}$), mouth +0.616 vs +0.334 (+0.282 [+0.168, +0.406], $p < 10^{-3}$).

The 10 unpaired rows, put back. A complete-case interval does not say what the missing rows could have done, and here they are missing for one reason: a narration that names no part cites the empty region. Restoring the empty region is the identity, so that narration's necessity is exactly 0 —

a property of the operator, not an imputed value — and the convention is applied on whichever side is empty. It is not a one-way rule: 5 rows are empty on the own side (which lowers the difference) and 5 on the shuffled side (which raises it). On all 150 rows the contrast is 0.3005 $[0.2347, 0.3683]$, against 0.3000 $[0.2342, 0.3691]$ on the 140 complete cases (Table 19). Assigning every missing row the worst value an individual difference of two probabilities can take bounds the full-cohort mean at 0.2133; the sign survives that bound (yes). These are deterministic bounds on the mean, not confidence intervals.

What the part-name parser is, and how much its coarseness could cost. The shipped parser is a *mention detector* over the forgery-reasoning field, not a semantic parser: a group counts as named when any of its surface words occurs. It therefore also fires on a part described as normal ("the smile and teeth look natural") and on a part used as a location anchor for an artifact in adjacent skin ("the skin around the eyebrows"). Across the 150 narrations it records 353 asserted mentions, 180 exculpatory ones and 125 anchors; Table 20 prints a stratified subset verbatim so the mapping can be checked without rerunning a model. A negation is *not* treated as a strip rule, because on these narrations a negation usually asserts the artifact ("does not match the rest of the face"); stripping negated mentions would delete valid citations. Because this corpus swaps exactly one group, every pixel outside it is already authentic and restoring it is the identity, so the intervention is governed by one bit — whether the cited union contains the swapped group. That identity holds exactly on 0.983 of the stored cells, with a maximum residual of 0.0488 where adjacent annotation masks overlap the swapped support; we report the residual rather than assume it away. A strict semantic reader can only *remove* mentions, so it can only flip that bit downwards: it names the swapped group on 0.84 of rows against the shipped parser's 0.96, flipping 18. Giving every flipped row the worst admissible difference still leaves the contrast at 0.1390 (sign survives: yes). The part masks themselves are the corpus's stored annotation masks, not a segmenter's output.

Sufficiency on the same corpus. Pasting the named parts *into* the authentic source raises the verdict by +0.635 against +0.352 for the parts a shuffled narration names (paired +0.299 [+0.233, +0.368], permutation $p < 10^{-4}$), with a ceiling of +0.641 and the identity gap at 0.0e+00; the named region is sufficient for 100.0% of the margin on the median (99.1% on the mean). Against the cheaper floor — a disjoint random window of the same shape, which on this corpus restores nothing of the swap — the gap is +0.687 [+0.618, +0.752] (permutation $p < 10^{-4}$, sign 109 of 111 positive). Necessity and sufficiency coincide here (Spearman +1.000) because the swap is exactly the named parts: on this corpus the two directions are one measurement. That is a fact about *this corpus geometry*, and it is the reason the FakeShield cohort — where the two directions came apart — cannot be read as the same measurement on a harder model. Which of the two explanations, corpus or system, produces that difference is not identified by these two runs.

What this licenses, and what it does not. On a corpus where the citation is the narration's own words and the control rules out a habit, the region FFAA names affects its verdict: the named-parts/full-region normalized intervention ratio is 0.992 [0.961, 1.000]. FakeShield's output-region ratio on the certified splice corpus is 0.2179 $[0.1158, 0.3289]$, and **the two numbers are not a ranking of the two models.** They are two audited configurations that differ in every layer between a released system and a number: a face corpus with one swapped anatomical group against a splice/copy-move corpus, a part lexicon against a native segmentation mask, a teacher-forced verdict field against an elicited first-token suffix, and different denominator-eligible cohorts (Table 21 lays the layers out). Separating intrinsic model faithfulness from measurement design would need a common task, region definition, readout validation and eligibility rule — an experiment this paper does not run. What the pair does establish is that the protocol returns *both* outcomes on released systems, so a null it reports is not an artefact of an instrument that can only return nulls. Three things remain open. The construction is a hard-paste region replacement, easier to detect than a modern swap (AUC 0.923 here against 0.997 on the eight-part corpus); the narration's part vocabulary is coarse (three groups), so a finer or wild-swap corpus could still find the narration citing a part the verdict does not use; and the shuffled control bounds only the habit explanation, not every alternative (a narration conditioned on the image through some cue other than the swapped pixels would pass it). Within those bounds the configurations exhibit different operational effects; which part of the difference is the system and which is the corpus, region instrument and readout is exactly what these two runs cannot separate.

Table 19: **THE FULL 150-ROW COHORT**: complete cases, the empty-citation identity applied to the 10 unpaired rows, and extreme-value bounds on the mean.

cohort	n	own − shuffled necessity [95% CI]
complete case	140	+0.3000 $[+0.2342, +0.3691]$
full cohort, empty citation $= 0$	150	+0.3005 $[+0.2347, +0.3683]$
full cohort, worst admissible	150	+0.2133 (bound)
full cohort, best admissible	150	+0.3467 (bound)

Table 20: **THE TWO CONSTRUCTIONS A MENTION DETECTOR CANNOT SEE**, verbatim from the narrations: counted by the shipped parser, dropped by the strict reader.

group	strict reader	sentence (verbatim from the narration)
eyes	exculpatory	The skin around the eyebrows seems overly smooth and lacks the natural texture found on the rest of the face.
eyes	exculpatory	The skin around the eyes and forehead appears overly smooth and lacks natural skin texture.
eyes	exculpatory	The skin texture around the eyes and forehead is unusually smooth and lacks natural pores and imperfections.
eyes	exculpatory	Additionally, the area around the eyes and the blending of the hairline with the forehead seem too perfect, lacking the natural variations typically seen in real photographs.
eyes	exculpatory	The facial features, including the eyes, nose, and mouth, are well-defined and consistent with natural human anatomy.
nose	exculpatory	The facial features, including the eyes, nose, and mouth, are well-defined and consistent with natural human anatomy.
eyes	landmark	Additionally, the area around the eyes and forehead shows inconsistent skin tones and blending issues, indicating possible tampering.
eyes	landmark	There are also inconsistencies in the skin color and texture, particularly around the eyes and forehead, suggesting manipulation.
eyes	landmark	Additionally, the area around the eyebrows looks slightly blurred and less detailed compared to the rest of the face, suggesting that the eyebrows were digitally altered or added.
eyes	landmark	Additionally, there is a slight blur and pixelation around the edges of the facial features, particularly around the eyes and eyebrows, indicating possible manipulation.
eyes	landmark	The area around the eyes and the eyebrows shows signs of manipulation, with an unnatural blending of skin tones.
eyes	landmark	Additionally, the area around the eyes and eyebrows shows signs of blending issues, with slight blurring and inconsistent skin tones.

Sufficiency: the direction the conclusion names. F Necessity measures the effect of restoring a region in the manipulated image; sufficiency measures the effect of inserting it into the authentic source. At the full ground-truth mask on a certified corpus, the two effects coincide: pasting the manipulated pixels of $M^\star$ *into* the authentic source is the manipulated image, so the sufficiency ceiling

Table 21: **THE LAYERS BETWEEN A RELEASED MODEL AND A NUMBER**, read from the audit manifests; paste-back is the operator; normalized n is the denominator-eligible cohort.

released system	native output	graded readout	region source	normalized n
FakeShield	free prose, no box	added yes/no suffix; first-token softmax *(elicited)*	MFLM, its own SAM localiser *(not narration)*	49 of 50
FFAA	released answer format, no box	its own verdict field, teacher-forced	annotation masks of the parts the narration names	133 of 150

equals the necessity ceiling, and we measure that identity rather than assume it (maximum gap between the pasted image's verdict and the stored $p(x)$: 0.0e+00 for FakeShield, 0.0e+00 for FFAA). For FakeShield on the $n = 150$ splice rows of Section E.3, pasting in only the cited region raises the verdict by +0.148 (median +0.1344) against +0.095 (median +0.0000) for a disjoint random window of the same shape: paired +0.053 [+0.019, +0.087], median +0.0131, sign 67 of 108 positive ($p = 0.016$), Wilcoxon $p = 0.001$, permutation $p = 0.002$, leave-two-out +0.0535. The cited region is therefore *sufficient* for an operational normalized effect of 56.6% on the median (51.2% on the mean, $n = 149$), while Section E.3 found a far smaller normalized necessity. That asymmetry is consistent with redundant or interacting cues; it is not an additive allocation of internal evidence. The two directions correlate across images (Spearman +0.355, $p < 10^{-4}$). For FFAA through the post-hoc grounder the random-window floor is not a floor: the swapped region is central and the grounded window peripheral, so a random translate lands on the edit more often than the citation does (+0.467 against +0.210); the narration-conditioned floor of the single-part corpus above is the one to read.

The other nine paraphrases. Read under each of the ten released paraphrases on every audited image and its source (3000 forward passes), the decision agrees with the audited paraphrase on 99.3%–100.0% of images, detection AUC ranges over 0.987–0.998, the per-image spread of $p(\text{fake})$ across paraphrases has median 0.001 and 90th percentile 0.005, and 0.7% of images change decision under any paraphrase. **This is a decision- and detection-level check on the audited images; it is not a replication of the intervention effects.** The narration-conditioned results above — the part-name citation test, its shuffled-narration control and the sufficiency direction — were run under paraphrase 0 only, and nothing here establishes that their magnitudes reproduce under the other nine. What it does establish is that the verdict this audit reads is not an artefact of the one paraphrase held fixed.

Cost. 4491 local forward passes and 150 greedy generations, 71 minutes on one accelerator (28.3 s per sample), zero network requests; `make ffaa` reproduces it, resuming from stored rows (Appendix N).

F THE SECONDARY CONTRAST: THREE GENERAL-PURPOSE CHAT VLMS

Section E audited the population this paper indicts. This section reports the control group it is paired against: frozen open-weight general-purpose VLMs served locally and *prompted* to do forensics, on the same 50 MagicBrush edits, each under a fixed 31-call ledger. Chat models prompted to do forensics are not the population this paper indicts, so they are a control group rather than the audit, kept for three reasons that neither released-forensic-VLM section could supply on its own.

1. **They rule out "the null is only a weak detector".** FakeShield's citation carries no measurable causal signal, and it is also a poor detector (AUC 0.587 $[0.520, 0.663]$), so one could ask whether the second fact explains the first. `qwen2.5vl:7b` is the strongest *served* model on this corpus (AUC 0.868 $[0.801, 0.928]$ against the runner-up's 0.726; several detectors elsewhere in this paper score higher), it detects on 30 of the same pairs rather than 8, and its ΔNS is null too. Detection strength does not buy citation faithfulness on this corpus.
2. **They bound the grounding confound.** These models emit their own boxes; FakeShield's explanation head emits none (0% of 50 rows), so its citation comes from our post-hoc grounder. Every citation-dependent comparison in Section E carries that asymmetry, and the served arm is where a reader can see what a self-grounded model's numbers look like under the same instrument.

3. **Their citations use boxes emitted by the answering model.** No grounder sits between narration and region. (They are not the only region a *system* outputs: FakeShield's MFLM mask is one too, Section E.6, but MFLM is a separate module of that system rather than the answering head, and it emits a mask, not a box.) On a corpus whose evidence is provably local the region `qwen2.5vl:7b` cites reaches a normalized intervention effect of 0.335 $[0.216, 0.472]$ (§I), measurably better than a control — on that hard-paste construction; the same ratio is not a construction-independent model property, because it is not estimable on the blended variant. It is not the paper's only positive citation result: FakeShield's own is 0.0233 $[0.0050, 0.0563]$ on the certified corpus, and FFAA's own-versus-shuffled narration contrast is +0.300 [+0.234, +0.369].

We present this arm as a contrast and a demonstration that the protocol runs end-to-end on served models, not as a survey of the field. The full treatment — prompts, the coverage repair, per-model tables and the correlation grid with its BH correction — is Appendix L.2.

First, a screen: which served models can do the task at all? A model that cannot detect has nothing to be faithful about, so before any audit we screen *every* vision model the shared server hosts on the same 50 pairs, at two calls per (model, pair), under a rule fixed in advance: keep a model iff the bootstrap interval on its paired AUC excludes 0.5 *and* a paired sign test on $p(x) - p(x_{\mathrm{src}})$ rejects at 0.05 (Table 22).

The screen does not single out one model. Under it, 5 of 5 served models pass — `gemma3:12b`, `llama3.2-vision:11b`, `llava:13b`, `llava:7b`, `qwen2.5vl:7b` — every one of them carrying statistically non-null *ranking* signal. What separates them is not whether the signal exists but whether the *thresholded verdict* is usable, which the flag-rate margin (flag rate on edits minus flag rate on authentic images) makes plain: `qwen2.5vl:7b` +0.60, `llama3.2-vision:11b` +0.24, `gemma3:12b` +0.20, `llava:13b` +0.06, `llava:7b` +0.00. `llava:7b` flags every image as manipulated, authentic ones included, so its detected stratum is empty and a faithfulness question is not well posed on it; `llava:13b`'s margin of $+0.06$ is barely better. Ranking signal and a usable thresholded verdict are distinct quantities.

What follows for the audit, stated as a gap rather than a finding. 2 models pass the screen and were *not* audited (`gemma3:12b`, `llama3.2-vision:11b`) — `llama3.2-vision:11b` with the second-best AUC in the whole screen (0.726) and a usable margin of $+0.24$. The reason is wall clock on a shared server where those two are not resident and are reloaded per request, which is an infrastructure cost and not a property of the models; it is recorded in `configs/real.yaml` and in `notes/NOTES.md`. So the honest summary of the served arm is: the protocol runs, the screen is cheap and should always be run first, one model was audited in depth, and two models that deserved the same treatment did not get it.

Reading Table 22. The rule is fixed in advance: keep a model iff the bootstrap interval on paired AUC excludes 0.5 and a paired sign test rejects. Every screened model passes; the "flag margin" column is why that is not the same as "every model can be audited".

What the contrast finds.

- **It points better than random while its verdict is not driven by what it points at.** On the 30 pairs `qwen2.5vl:7b` detects it reaches 0.129 $[0.086, 0.183]$ against an unfaithful-citation null of 0.052; on its rows $\rho(\mathrm{CF}, \Delta\mathrm{NS}) = +0.373$[+0.054, +0.623] against $\rho(\mathrm{CF}, \mathrm{IoU}) = +0.937$[+0.868, +0.971] — the correlation that motivates separate reporting of the factors (Appendix O).
- **On a corpus where the evidence is provably local** (§I), restoring the true mask recovers the entire detection margin — necessity at $M^\star$ rises from 0.1236 $[0.0766, 0.1836]$ to 0.7250 $[0.6337, 0.7938]$ — and restoring the region the model cites reaches a normalized intervention effect of 0.335 $[0.216, 0.472]$. The *ceiling* was corpus-induced; the *citation* was not.
- **Every causal statement about served models in this paper is $n = 1$ model**, and we say so here rather than in a subsection — not because only one model can detect (the screen shows all 5 carry signal) but because only one was audited with a usable verdict margin. That is a limitation of compute, and the screen is what makes it visible instead of inferable.

Table 22: Detection screen across every served vision model, 50 paired MagicBrush triples, two calls per (model, pair).

Model	n	paired AUC	sign +/-	sign p	detects?	flag rate (edit)	flag rate (authentic)	flag margin	audited?
qwen2.5vl:7b	50	0.868 [0.801, 0.928]	46/4	0.0000	yes	0.80	0.20	+0.60	yes
llama3.2-vision:11b	50	0.726 [0.663, 0.798]	44/6	0.0000	yes	0.92	0.68	+0.24	no (wall clock)
gemma3:12b	50	0.655 [0.570, 0.738]	34/16	0.0153	yes	0.36	0.16	+0.20	no (wall clock)
llava:7b	50	0.627 [0.576, 0.688]	38/12	0.0003	yes	1.00	1.00	+0.00	yes
llava:13b	50	0.567 [0.521, 0.619]	33/17	0.0328	yes	0.72	0.66	+0.06	yes

EDCT's explanation-change leg. CCS has two legs. Its answer leg is recomputed from stored probabilities at zero query cost, which is the comparison the positive control makes (§3.2); its explanation leg we buy on the served rows — one extra generation per row, re-explaining $A_{R(x)}(x)$ and applying EDCT's own two triggers (ROUGE-L similarity below 0.9, or re-grounded region IoU below 0.5) — and report CCS's *disjunction* beside its answer leg (Table 23). Measured: `llava:13b` answer 0.080, explanation 0.480, disjunction 0.500; `llava:7b` answer 0.000, explanation 0.380, disjunction 0.380; `qwen2.5vl:7b` answer 0.120, explanation 0.640, disjunction 0.660. The disjunction is carried by the *explanation* leg, not the answer leg — on `qwen2.5vl:7b` the answer leg fires on 0.120 of rows and the explanation leg on 0.640, so EDCT's criterion credits a model that never changes its verdict with 0.660 consistency purely because its narration drifts under an intervention that is the identity map on authentic content. That is a property of the criterion, not of our reimplementation, and it is the sharpest reason we report necessity as a graded quantity rather than a flip test. The similarity threshold is *ours* — EDCT decides this leg with LLM-assisted analysis — so we sweep it and report the whole curve rather than one number: on `qwen2.5vl:7b` the disjunction reads 0.24 at $\tau = 0.3$, 0.42 at $\tau = 0.5$, 0.60 at $\tau = 0.7$, 0.66 at $\tau = 0.9$. It never falls to the answer leg's 0.120 at any threshold we tried, which is the point: the criterion's value is set by how much narration drift one is willing to call an explanation change.

Reading Table 23. The disjunction is EDCT's criterion; the answer leg alone is the zero-cost comparison the positive control makes.

G Sensitivity to the counterfactual operator

Explanation-faithfulness work outside forensics has no choice but to *synthesise* its counterfactual: erase a region and inpaint it, blur it, replace it with a generative fill. Manipulation forensics is the rare domain where the counterfactual is *held* rather than synthesised — the authentic source of the edit exists — and this section shows that the difference is not a convenience. It changes the sign of the answer.

The confound, stated precisely. Let A_R be the restoration operator. The quantity we want is $\mathrm{Nec} = p(x) - p(A_R(x))$, and it is only a measure of *evidence removal* if A_R writes pixels the detector regards as authentic. An inpainter — classical or learned — writes pixels the detector did not see in any authentic image, so it removes one manipulation and inserts another, and Nec becomes a difference of two effects with opposite signs and no way to separate them post hoc. Paste-back has no such term: restoring authentic pixels into an authentic image is the *identity map*, so its false-flip rate on authentic content is exactly 0 at zero query cost, and this is asserted in the test suite rather than measured (Proposition 2).

Measured on served models: the operator inverts the sign of necessity. Pooled over 150 served-model rows, necessity is +0.001 $[-0.020, +0.023]$ under paste-back, -0.140 $[-0.175, -0.110]$ under the Laplace fill and -0.140 $[-0.174, -0.110]$ under the noise-matched fill (both $p < 10^{-4}$ against gold, paired sign-flip). Rank agreement with the gold operator is weak ($\rho = +0.12$ $[-0.04, +0.28]$ and $\rho = +0.19$ $[+0.03, +0.34]$), so not even the *ordering* of examples survives the substitution. The false-flip calibration says why: on authentic images the two inpainters push 31.3% $[24.5 - -39.1\%]$ (47 of 150) of them across the decision boundary, with mean effects of +0.273 and +0.256 in probability units — larger than the necessity those same served models show at the *full* edited region on this corpus (0.1236 $[0.0766, 0.1836]$), so on this arm the operator's own trace exceeds the signal it was meant to measure. It is not larger than every necessity in this paper: the certified corpora and the trained arms reach several times that (Table 4).

Table 23: EDCT's Counterfactual Consistency Score with *both* legs, computed with our operator on the identical served rows, images and citations.

Model	n	CCS answer leg	CCS explanation leg	CCS disjunction	ROUGE-L(expl, expl of CF)
`llava:13b`	50	0.080	0.480	0.500	0.802
`llava:7b`	50	0.000	0.380	0.380	0.847
`qwen2.5vl:7b`	50	0.120	0.640	0.660	0.710

Accuracy against a known oracle effect. The served-model comparison shows the operators disagree; it cannot show which is right, because the true necessity of a served VLM is not known. The positive control supplies exactly that missing ingredient: on the oracle-faithful detector of Section D, applied to 100 of the same MagicBrush triples, the true necessity is fixed by construction. The classical inpainters recover the template-matching oracle's effect closely. Paste-back recovers 0.991 $[0.991, 0.992]$; the Laplace fill reports 0.978 $[0.966, 0.983]$ (error -0.013) and the noise-matched fill 0.982 $[0.975, 0.986]$ (error -0.009), and both have a false-flip rate of 0.00 on authentic images — against 31.3% on the served VLMs.

Two more oracle families, and they disagree with the first — which is the point. A conclusion drawn from one detector of known necessity is a conclusion about that detector. We therefore ran the same two operators, on the same 100 triples, against two further families chosen for the mechanism they key on.

The blind anomaly detector scores the mean high-pass noise energy of the block it cites — i.e. it is exactly the family that *should* be trace-sensitive. Its true necessity is -0.067 $[-0.086, -0.053]$ (negative: restoring authentic pixels there *raises* its anomaly reading). The Laplace fill reports 0.031 $[0.024, 0.050]$ and the noise-matched fill 0.027 $[0.019, 0.044]$, errors of +0.098 and +0.094 — **a sign inversion**, on the same images, with the same operators that were accurate to -0.013 on the template-match oracle. That is the detector-dependence claim demonstrated within the positive control rather than inferred across it: the operator error here spans an order of magnitude and both signs depending only on *which detector* is audited.

The global-cue oracle reads pixels *outside* the region the operator is handed, which makes its false-flip rate a direct probe of whether the operator respects the region it was given. Under paste-back and the Laplace fill it flips 0.000 of authentic images; under the noise-matched fill it flips **0.790**, with a mean shift of +0.616 in probability units. The cause is mechanical and we verified it directly rather than guessing (Table 25): the noise-matched operator adds noise across the canvas and composites with a *feathered* mask, so on a real MagicBrush citation region it changes a mean of 1944 pixels outside the nominal region — 0.076 of its own edit — with a maximum magnitude of 0.0856. Paste-back and the Laplace fill change *exactly zero* pixels outside, on all 40 images tested.

This implementation-specific containment failure illustrates a general risk: any counterfactual metric attributes necessity to the region it hands the operator, and a soft-boundary operator silently violates that attribution while every quantity in the ledger continues to look reasonable. We ship the screen (`rightpixels.counterfactual.base.region_containment`) and a test that pins it, and we recommend running it on any new operator before its numbers are believed. The served audit's 31.3% false-flip figure uses this operator, so its containment failure is a limitation of that comparison.

The contrast reflects detector sensitivity. The oracle scores whether *this specific edit* is still present; it is by construction blind to generative traces, so an operator that injects fresh ones costs it almost nothing. The served VLMs are the opposite: a large part of what makes them say MANIPULATED is a generative fingerprint, so an inpainter hands them the very cue they are keyed on. **The size of the inpainting confound is therefore a property of the detector under audit, not of inpainting in the abstract**, and a universal claim that inpainting is invalid is not supported. What the data support is sharper and more useful: the confound ranges from negligible to sign-inverting across detectors evaluated on the *same* images with the *same* operators, so it cannot be assumed either way and *must* be measured per detector. The false-flip calibration measures it for one extra forward pass per operator per image.

Table 24: **OPERATOR ERROR AGAINST A KNOWN GROUND TRUTH, ACROSS THREE DETECTOR FAMILIES.** The same restoration operators applied to detectors whose true necessity is fixed by construction ($n = 100$ MagicBrush triples each).

Detector	Operator	measured Nec	true Nec	error	false-flip rate	mean dp authentic
ORACLE-FAITHFUL (alpha=1)	paste-back	0.991 [0.991, 0.992]	0.991	+0.000	0.000	0.000
ORACLE-FAITHFUL (alpha=1)	Laplace fill	0.978 [0.966, 0.983]	0.991	−0.013	0.000	0.011
ORACLE-FAITHFUL (alpha=1)	noise-matched fill	0.982 [0.975, 0.986]	0.991	−0.009	0.000	0.007
partial alpha=0.50	paste-back	0.180 [0.147, 0.222]	0.180	+0.000	0.000	0.000
partial alpha=0.50	Laplace fill	0.136 [0.107, 0.173]	0.180	−0.044	0.000	0.001
partial alpha=0.50	noise-matched fill	0.147 [0.117, 0.186]	0.180	−0.033	0.000	0.001
ORACLE-UNFAITHFUL (decoy)	paste-back	0.000 [0.000, 0.001]	0.000	+0.000	0.000	0.000
ORACLE-UNFAITHFUL (decoy)	Laplace fill	0.000 [0.000, 0.001]	0.000	−0.000	0.000	0.000
ORACLE-UNFAITHFUL (decoy)	noise-matched fill	0.000 [0.000, 0.001]	0.000	−0.000	0.000	0.000
blind anomaly detector	paste-back	−0.067 [−0.086, −0.053]	−0.067	+0.000	0.000	0.000
blind anomaly detector	Laplace fill	0.031 [0.024, 0.050]	−0.067	+0.098	0.000	−0.052
blind anomaly detector	noise-matched fill	0.027 [0.019, 0.044]	−0.067	+0.094	0.000	−0.045
GLOBAL-CUE (true box, non-causal verdict)	paste-back	0.000 [0.000, 0.000]	0.000	+0.000	0.000	0.000
GLOBAL-CUE (true box, non-causal verdict)	Laplace fill	−0.000 [−0.000, −0.000]	0.000	−0.000	0.000	0.000
GLOBAL-CUE (true box, non-causal verdict)	noise-matched fill	−0.000 [−0.002, −0.000]	0.000	−0.000	0.790	0.616

Reading Table 24. "error" is the operator's estimate minus the truth. Read down the *Detector* column: the template-match oracles are recovered to ≤ 0.05, the noise-keyed `blind_anomaly` detector is recovered with the *wrong sign*, and the global-cue detector — which reads pixels outside the region the operator is handed — is unaffected by two operators and flipped on 79% of authentic images by the third. Same images, same operators.

Reading Table 25. Does the operator keep its edit inside the region it was given? Measured on 40 real MagicBrush citation regions. A counterfactual metric attributes necessity to the region it hands the operator, and a control region is chosen to be disjoint from it; an operator that writes outside R breaks both, invisibly. Paste-back is the identity outside R by construction; the Laplace fill only touches the hole; the noise-matched fill composites through a feathered mask and leaks.

Reading Figure 9. The same operators are accurate to ≤ 0.05 on the template-match oracles and *sign-inverted* on the noise-keyed `blind_anomaly` detector. Right: false-flip rate on images that were never edited. Paste-back is the identity map there, so it is exactly zero at zero query cost; the noise-matched fill flips 79% of them for the global-cue detector, because that detector reads the pixels the operator leaks into (Table 25).

What this means for prior work, and for anyone who reuses this protocol. Counterfactual explanation evaluations built on *synthesised* restoration are, for detectors that respond to the traces the operator writes, partly measuring their operator — and modern forensic VLMs are exactly such detectors, which is the whole reason they work at all. We do *not* claim the effect is universal; the oracle above is the counterexample we found in our own data and report.

The learned inpainter, measured. The extension from our two classical numpy fills to a *learned* inpainter is a measurement rather than a conjecture: we ran **big-lama** (Suvorov et al., 2022), the released torchscript checkpoint, through the identical operator study on 100 MagicBrush triples and four oracle families whose true necessity is known.

- **It does not recover a known necessity.** On the faithful oracle, whose true necessity under paste-back is 0.9913, LaMa reports 0.9449 $[0.9068, 0.9632]$ — an error of -0.0464, against -0.0126 for the Laplace fill and -0.0086 for the noise-matched one. On the half-evidence oracle (true 0.9915) the error grows to -0.1079. A *learned* operator is thus 4–$12\times$ further from the known answer than the hand-written ones, which is the opposite of what a reader would assume from "more realistic fill".
- **Its false-flip rate is not zero.** On authentic images LaMa moves the faithful oracle by +0.0404 and pushes 2.0% of them across the decision boundary, where paste-back is the identity map and both numbers are exactly 0 by construction.
- **The sign inversion reproduces.** On the noise-keyed `blind_anomaly` detector the true effect is -0.0670 and LaMa reports 0.0295 $[0.0221, 0.0470]$ — the same sign flip the classical fills produce, so the detector-dependence result is a property of *restoration by synthesis*, not of numpy.

Table 25: **OPERATOR REGION-CONTAINMENT SCREEN.**

Operator	images	changed pixels outside R (mean)	fraction of the edit outside R	max |change| outside R	contained on every image
paste-back	40	0	0.000	0.0000	yes
Laplace fill	40	0	0.000	0.0000	yes
noise-matched fill	40	1944	0.076	0.0856	NO

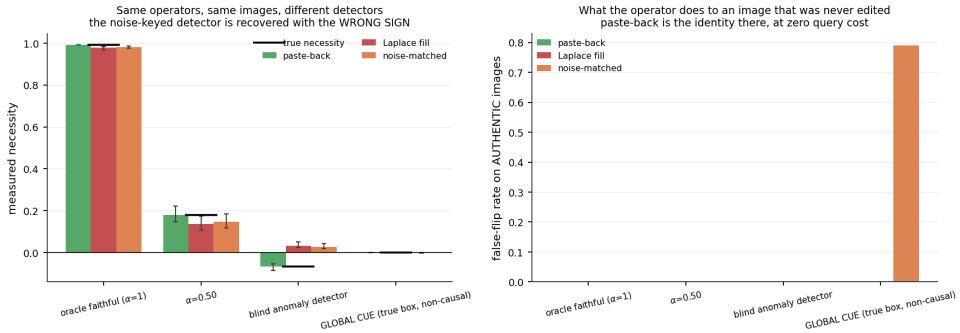


Figure 9: **The operator is not a detail, and the confound belongs to the detector.** Left: necessity as reported by three restoration operators for four detectors whose true necessity is known (black ticks).

- **Region containment.** LaMa is perfectly region-contained: on the global-cue oracle, which reads only pixels *outside* the region, LaMa's necessity is -0.0000 $[-0.0000, -0.0000]$ and its false-flip rate 0.0%, while *our own* noise-matched fill leaks (+0.6159 mean effect, 79.0% of authentic images flipped). The containment bug found in Table 25 is ours, not inpainting's.

We therefore state the operator result without a hedge: on detectors that respond to synthesis traces, a synthesised counterfactual measures the operator as much as the model, and the error against a known necessity is *larger* with a learned inpainter than with a classical one (the region-containment bullet above is the one measurement that runs the other way, and we report it). Choosing an inpainter alone provides no small-error guarantee. Its effect is larger than the local-restoration signal on the three evaluated served models, and it is *measurable with an additional forward pass* by running the same operator on authentic images. We therefore make three recommendations, all of which this paper follows:

1. **Prefer a held counterfactual to a synthesised one** whenever the domain supplies it. It has an exact authentic-input identity and introduces no new synthesis trace; partial-restoration interactions still require evaluation. Forensics supplies it; so do domains with paired pre/post data (medical follow-up imaging, before/after edits, A/B interventions).
2. **Publish the false-flip rate next to every faithfulness number.** It is one extra forward pass per operator per image, on data you already have, and without it a faithfulness score has no error bar in the only direction that matters.
3. **Report the operator in the metric's name**, not in the appendix. "Necessity" under paste-back and "necessity" under LaMa are different quantities that disagree in sign, and calling them by one name is how the disagreement stayed invisible. This applies to us: the operator we ship under the registry key `inpaint_diffusion` solves a heat equation, not a denoising-diffusion model. The key is retained because it is the key of every cached response and of every shipped results file.

H INSIDE A REAL VLM: ACTIVATION PATCHING AND FAITHALIGN

Section L.2 is black-box by necessity: a serving API exposes logprobs and text and nothing else. Two of this paper's claims cannot be settled that way. Whether the narration is produced by the same evidence pathway as the verdict is a statement about internal computation, and whether CF-Score can be *optimised* is a statement about gradients. We therefore run both on a local, open-weight VLM

— **Qwen2-VL-2B-Instruct** (Wang et al., 2024), 2.21B parameters, float32 on Apple MPS — where the activations and the gradients are ours.

The same audit code runs against it: the white-box adapter duck-types onto the served one, so the CF-Score ledger, the grounder, the control matcher and the paste-back operator are byte-identical to Section L.2. One estimator improves: holding the full logit vector, the verdict renormalisation sums over the *entire* vocabulary (330 manipulated-family and 205 authentic-family token ids) rather than a top-20 truncation, so `mass_outside` is exact.

The starting point is a weak, mostly always-fake responder. Untrained, Qwen2-VL-2B answers MANIPULATED to nearly everything: across the held-out splits its detection AUC is 0.746 ± 0.020 (per split 0.760, 0.732), its IoU against the true edit is 0.131 ± 0.012 and its CF-Score is 0.0460 ± 0.0121. It is, in other words, close to the same failure mode as `llava:7b` at a quarter of the size. That makes it an honest "before" state rather than a flattering one — and the per-split spread is itself a reason to report splits: how weak the starting point is depends on which thirty images it is evaluated on.

H.1 FaithAlign: what a counterfactual training term buys

We test whether counterfactual supervision changes restoration effects beyond verdict and explanation-imitation losses. We train three strictly nested LoRA arms on the same checkpoint, the same 80-triple training split, the same 150 steps and the same seed, and audit each on an edit-ID-disjoint 30-triple held-out split:

- `detect`: L_{det}, verdict cross-entropy on (edit, source).
- `detect_cite`: $+ L_{cite}$, token cross-entropy on a reference explanation that names and boxes the true edit — this is what "train the model to explain" means in the literature, imitation of a good rationale.
- `faithalign`: $+ \lambda L_{cf}$, where L_{cf} requires the *restoration effect* to be localised on the true region:

$$L_{cf} = -\log \frac{\exp(e_0/T)}{\sum_k \exp(e_k/T)}, \qquad e_k = p(x) - p(A_{R_k}(x)),$$

with $R_0 = M^\star$ and $R_1, \ldots, R_K$ congruent disjoint controls.

Softmax and hinge counterfactual objectives. Eq. 15 defines $\mathcal{L}_{\text{cf}}$ as a *margin hinge* on the ground-truth restoration effect plus a hinge *penalising* the control effect. The arm named `faithalign` above trains a *softmax* over restoration effects instead. A reader following Eq. 15 would expect the control hinge to have prevented exactly the failure we report — and would be right: a softmax is invariant to adding a constant to every effect, so raising every restoration effect equally leaves this loss unchanged, whereas the second hinge of Eq. 15 does not. We also train the hinge in Eq. 15 as a fourth arm, `faithalign_hinge` ($m_c = 0.5$, $\epsilon_c = 0.05$), and report it beside the softmax arm in Table 28. Two further terms of Eq. 14 — $\mathcal{L}_{\text{attr}}$ and $\mathcal{L}_{\text{size}}$ — are **not** trained on any real VLM here, and the limitations section says so rather than leaving the Method's five-term objective to imply otherwise.

L_{cf} is *not* a restatement of L_{det}. A detector can ace edit-versus-source classification using a global cue — DALL-E 2 re-encodes the whole canvas — while every regional restoration leaves the verdict untouched, which is precisely the served-model failure this paper measures, where restoring even the true mask moves `qwen2.5vl:7b` by only 0.124 of a total separation of 0.453. The `detect` arm exists so that "the gain is just better detection" is a hypothesis we test rather than a caveat we write.

Necessity under the nested training objectives. Held-out necessity is effectively zero for the untrained checkpoint and small for the `detect` arm: across splits `detect` reaches AUC 0.992 $\pm$ 0.011 with necessity 0.029 ± 0.041 (per split 0.058, 0.000). A model that separates edits from photographs almost perfectly therefore barely moves when the region it cites is restored — one to two orders of magnitude below the 0.99 a detector of *known* faithfulness reaches on the same corpus

(Section D). That is the paper's central pathology, reproduced inside a model we trained ourselves, and it is the cleanest available refutation of "a good detector is automatically a faithful one".

The exact-zero necessity at AUC 1.000 occurs on one split only. Across permuted splits, the supported result is small, variable necessity, not a universal zero.

Imitating a good explanation buys some of it back, and adding the counterfactual term raises necessity further: `faithalign` reaches 0.261 ± 0.009 against `detect_cite`'s 0.059 ± 0.053, a paired gain of +0.202 $\pm$ 0.063 (positive in 2/2 splits, $p < 0.05$ within 2/2). Detection is essentially unchanged across the trained arms (`detect` 0.992 ± 0.011, `detect_cite` 0.992 ± 0.011, `faithalign` 0.984 ± 0.008), so "the gain is just better detection" — the alternative the `detect` arm exists to test — is ruled out.

The controlled contrast remains unresolved. Interpretation depends on control coverage.

Against a *congruent* control alone, the ΔNS and CF contrasts rest on almost nothing. On these arms that control barely exists: they cite very large regions, and the coverage audit of Section L.2 finds a congruent control on only 0.16 of the 55 held-out rows for `faithalign` and 0.29 for `detect`. The congruent-only CF contrast therefore rests on nine images and is underpowered. Under the coverage-complete control it rests on all 55, and it has a narrower interval that still includes zero: CF 0.0039 $[-0.0161, 0.0248]$ ($p = 0.7043$, $n = 55$) and ΔNS 0.0127 $[-0.0911, 0.1203]$ ($p = 0.8119$, $n = 55$) for `faithalign` over `detect_cite`.

The reason is visible once specificity is measurable on every row. `faithalign`'s necessity at its own citation is 0.2680[0.1920, 0.3609] — and its *specificity*, the effect of restoring an equally suspicious region it did *not* cite, is 0.2837[0.2005, 0.3849]. The paired increases over `detect_cite` are 0.2465[0.1603, 0.3342](p=0.0001,n=55) in necessity and 0.2197[0.1241, 0.3198](p=0.0001,n=55) in specificity: the same quantity, to within their intervals. L_{cf} did not teach the model *where* its verdict should depend on the image. It taught the model to respond to gold paste-back *anywhere*, which raises both legs of ΔNS together and leaves the contrast where it started. Localisation does not move either (0.0114 $[-0.0274, 0.0516]$ ($p = 0.5762$, $n = 55$)), so the product is unchanged for two independent reasons.

These results do not establish improved citation faithfulness. A counterfactual term defined as a softmax over *restoration effects* is satisfied by any weight assignment that makes the true region's effect largest; nothing in it penalises making every region's effect large. Constraining the control legs is the obvious next objective, it is represented by Eq. 15; the `faithalign_hinge` arm below trains that form.

The fixed-region contrast is positive. Necessity at the *fixed* ground-truth mask is the same region, operator and image set for every arm, so no arm can inflate it by citing differently and no control region is involved at all. There `faithalign` reaches 0.3916[0.2984, 0.4935] against `detect_cite`'s 0.0595[0.0065, 0.1492], a paired gain of 0.3320[0.2453, 0.4231](p=0.0001,n=55) and 0.3866[0.2951, 0.4827](p=0.0001,n=55) over the untrained checkpoint, on all 55 held-out triples with no rows dropped. That is the claim we make: the counterfactual term increases how much of a 2B model's verdict is removable by restoring the true edited pixels. It is *not* the claim that the model's own narration became more faithful; control coverage distinguishes these questions.

Replication across genuinely permuted train/eval splits. Replication requires distinct evaluation splits, not just different random seeds. The runner permutes explicitly, writes an `eval_split_fingerprint` into every manifest, and rejects identical evaluation fingerprints as separate split replications.

We report 2 splits (eval fingerprints `621f9b4f6924`, `7b38b46283c2`; largest pairwise overlap 13 of 55 held-out images), and we are explicit about their provenance because it is not uniform: one is the original deterministic split — included once in the split-level summary — and the others are freshly permuted and additionally audited under the two classical inpainters. The split is the unit of replication throughout, so every number below is a mean over splits with the across-split spread beside it, never a pooled interval that would treat repeated evaluations of overlapping image sets as independent. Across those splits the constructive result is:

- held-out necessity at the model's freely-chosen citation, per split: `detect_cite` 0.022, 0.097 versus `faithalign` 0.268, 0.255. The paired gain is +0.202 $\pm$ 0.063 (mean of per-split paired means $\pm$ across-split sd), positive in 2/2 splits and significant at $p < 0.05$ within 2/2;
- CF-Score does *not* follow: -0.037 ± 0.014, positive in 0/2 splits, and localisation gets slightly worse (-0.022 ± 0.048, positive in 1/2). Both replicate as negatives.

Fixed-region restoration effects across splits. Necessity at the model's own citation confounds *how much the verdict moves* with *what region the model chose*: an arm that cites a larger or better-placed region gets a larger necessity for free. The audit ledger has always also recorded `necessity_ceiling_gt` — the necessity of restoring the *entire ground-truth mask* — which is the same region, the same operator and the same images for *every* arm, and is therefore area- and region-matched by construction. The per-split estimates are: `detect_cite` 0.060, 0.033 against `faithalign` 0.392, 0.155, a paired gain of +0.227 $\pm$ 0.149, positive in 2/2 splits. This comparison supports a narrower training result: **the counterfactual term makes the verdict depend on the manipulated pixels, measured on a fixed region that no arm chose**.

H.2 The hinge objective and the controlled contrast

The null above is a property of the *softmax* form, not of the objective the Method section prints. We therefore train Eq. 15 literally — a margin hinge on the ground-truth restoration effect ($m_c = 0.5$) plus a hinge penalising the control effect ($\epsilon_c = 0.05$) — as `faithalign_hinge`, on the same split, steps, seed and data as `detect_cite`, and audit all 40 held-out triples under the coverage-complete control.

It moves the quantity the softmax could not. Paired against `detect_cite` on 40 rows: ΔNS rises by +0.1278 $[+0.0403, +0.2317]$ ($p = 0.004$) — against the softmax arm's 0.0127 $[-0.0911, 0.1203]$ ($p = 0.8119$, $n = 55$), a null we replicated three times. And it gets there the way Eq. 15 says it should: necessity at the citation does *not* rise (-0.0188 $[-0.0658, +0.0085]$, $p = 0.718$), so the gain is carried almost entirely by the control leg falling — the second hinge doing exactly the job the softmax's shift-invariance made impossible. Detection is essentially unchanged (AUC 0.962 against 0.973).

We report two things that did not follow. Necessity at the fixed ground-truth mask rises by +0.0613 $[-0.0007, +0.1361]$ but its interval touches zero ($p = 0.103$), so we do not claim it. And **CF-Score still does not move** (-0.0112 $[-0.0475, +0.0230]$, $p = 0.550$): the composite absorbs a real, significant causal gain because localisation drifted down by -0.0348 $[-0.1040, +0.0325]$ at the same time. That is the third independent occasion in this paper on which the product hides a movement in its own causal factor, supporting separate reporting of necessity and localisation.

H.3 The global-cue contrast, reproduced on a trained network

Every demonstration so far that the causal factor sees something localisation cannot has come from a detector we wrote — the global-cue oracle, whose necessity is 0 by algebra. That is a fair objection and the fix is to build the same cell out of a network. We train a fourth arm, `detect_cite_globalcue`, identical to `detect_cite` in checkpoint, steps, seed, data and *citation supervision*, differing in one thing: its **positive class is** $A_{M^\star}(x)$, the manipulated image with the edit region restored from the authentic source. Every locally decidable difference between its two classes has been deleted, so the only signal left is DALL-E 2's canvas-wide re-encode fingerprint. It learns it: held-out AUC 0.984 against `detect_cite`'s 0.973 on the same 40 triples.

The result, on 40 paired held-out rows. The two arms are matched where the design says they should be and separated where it says they should be:

paired contrast (`globalcue` − `detect_cite`)	value	p
localisation IoU	-0.0150 $[-0.0757, +0.0437]$	$= 0.647$
necessity at the cited region	$\mathbf{-0.0482}$ $[-0.1209, -0.0002]$	$\mathbf{= 0.011}$
CF-Score (coverage-complete)	-0.0133 $[-0.0465, +0.0189]$	$= 0.459$

Necessity separates two real trained networks that IoU cannot, and the composite does not. The global-cue arm's necessity at its own citation is -0.0000 $[-0.0001, 0.0001]$ — its interval includes zero, as the construction predicts — against 0.0481 $[0.0004, 0.1655]$ for the arm trained on the real target. Their citation-IoU contrast also has an interval containing zero; equivalence is not established. This is the cell the paper needed; the positive control exhibits it only on an oracle.

It also, reported because it is the more important half, **fails for the product**. CF-Score does not separate the two arms ($p = 0.459$): with $\Delta\mathrm{NS} \approx 0$ on both, the sigmoid sits at $\frac{1}{2}$ and the composite is $\frac{1}{2}\times$IoU, so it inherits IoU's null and discards the one factor that worked. The demonstrated added value of this paper's causal machinery on a real trained model therefore belongs to **necessity under paste-back**, reported beside IoU — not to $\mathrm{CF} = \sigma(\Delta\mathrm{NS}/T) \cdot \mathrm{IoU}$. The two quantities are reported separately throughout the main text.

H.3.1 Outlier, scale and multiplicity sensitivity

The paired distribution, probability scale and multiplicity family each qualify the mean contrast.

(a) The mean was carried by two of forty rows. Recomputing the paired differences from the stored rows: mean -0.0482 $[-0.1209, -0.0002]$, **paired median -0.0001**, sign 28 of 40 negative ($p = 0.0166$), and the mean with the two largest |differences| removed is -0.0004 — about 1% of the reported effect. Figure 10 plots all forty. Because two observations account for approximately 99% of the mean, its magnitude does not describe a typical image.

(b) Both arms were pinned at the top of the sigmoid. $p(x)$ has mean 0.9985 (min 0.9549) for `detect_cite` and 0.9997 (min 0.9985) for the global-cue arm; 98% and 100% of rows respectively sit above 0.99. A difference of probabilities therefore has $\sim 10^{-4}$ of range on almost every row, *whatever* pathway produces the verdict — so this saturated probability scale cannot separate "the global-cue arm has zero necessity" from "the arm has no remaining probability range." Re-expressing the identical stored quantities on the log-odds scale (Eq. 8) restores the range: mean -0.942 $[-1.918, -0.274]$, median -0.278, sign 29 of 40 negative ($p = 0.0064$), leave-two-out -0.414. On this scale 29 of 40 rows move in the predicted direction, the median is two orders of magnitude away from zero relative to its own scale, and 44% of the mean survives deleting the two largest rows. The same re-analysis applied to the ceiling gives -0.873 $[-1.249, -0.529]$ ($p < 10^{-4}$).

(c) Sensitivity to the multiplicity family. We report Benjamini–Hochberg corrections under three explicit family definitions. The log-odds analysis is post hoc (Table 5), not a preregistered endpoint:

family	size	q(raw necessity)	q(log-odds necessity)
initial contrasts only (the prespecified family)	18	≥ 0.0756	not in family
expanded white-box contrasts	35	0.0406	0.0125
strict union of both	53	0.0512 (does not survive)	**0.0189** (survives)

Under the strictest reading available — the union of all reported contrasts, 53 tests — the raw-probability contrast **does not survive** ($q = 0.0512$) and the log-odds contrast **survives** ($q = 0.0189$). This is consistent with the saturation diagnosis in (b), but does not turn the post hoc log-odds result into confirmatory evidence. The hinge arm's $\Delta\mathrm{NS}$ result survives as well ($q = 0.0247$), and the 9 of 53 contrasts that reject are listed in Table 26. We note that BH assumes independence or positive regression dependence, and that the raw and log-odds versions of one contrast are monotone re-expressions of the same stored probabilities — so the family size should not be read as 53 independent questions.

(c′) Log-odds effects are not uniformly more robust. The table above is the paper's strongest argument for reading necessity in log-odds, and it would be easy to leave the impression that the log-odds scale is the more robust one in general. It is not, and the cleanest evidence against that reading is our own: on the hybrid corpus of Section J, where the same two arms are replicated across 3 seeds with freshly permuted splits, the *log-odds* re-expression of the self-cited necessity contrast has per-seed means 0.404, -0.819, -0.358 (mean -0.258, across-seed sd 0.618), per-seed

permutation p 0.4841, $< 10^{-4}$, 0.2075, sign agreement 67%, every per-seed p below 0.05: **no — 1 of 3 seeds inverts the sign**. The region-matched contrast that section leads with is stable on the raw scale over the identical runs (-0.2764, -0.2772, -0.3206, sign agreement 100%, every per-seed p below 0.05: yes). So the argument for log-odds is narrow and mechanical — it is the scale that has range where the verdict is saturated (§C.2), which is the condition (b) establishes for *this* cell — and it is not an argument that log-odds numbers are more reliable. Where a contrast is not saturated, the log-odds transform amplifies exactly the rows sitting near 0 or 1, and the seed spread above is what that costs. A reader who wants one rule: read the scale that has range, and read the paired median and the sign test on both.

(d) One seed is still one seed. Everything above is a re-analysis; it cannot fix the fact that this cell is one training seed on one split at $n = 40$. Section J is where that is addressed: the same contrast on a corpus where necessity has real dynamic range, with all completed seed/split replications reported separately.

Reading Figure 10. Differences are sorted, on the raw-probability scale (left) and the log-odds scale (right), with the mean, the paired median and the leave-two-out mean marked. The left panel makes the concentration of the raw mean in two rows visible.

A control for the fixed ground-truth region. The objection is exact: $\mathrm{Nec}(M^\star)$ is the same *region* for every arm, but that is not the same as being *controlled*. Two paragraphs above we conclude that L_{cf} taught the model to respond to gold paste-back *anywhere*; if that is true, $\mathrm{Nec}(M^\star)$ rises for free and the constructive result cannot be separated from that alternative. The audit ledger therefore includes a 31st call: an **area-exact control for** $M^\star$ — a region of the same size as the ground-truth mask, placed elsewhere, restored with the same operator — giving

$$\Delta\mathrm{NS}(M^\star) \;=\; \mathrm{Nec}(M^\star) \;-\; \mathrm{Spec}(M^\star\text{-area-matched control}),$$

which is the arithmetic the region-matched leg alone omits. Section I reports it on both corpora (Table 31); on the local-evidence corpus it is exactly equal to $\mathrm{Nec}(M^\star)$ on every row, because restoring a region disjoint from $M^\star$ there rewrites no pixel at all, so the control cannot fail and the ceiling claim needs no statistical adjustment.

A non-monotonicity we have to report. On 38/85 of the `faithalign` rows, restoring the cited region moves the verdict *more* than restoring the whole ground-truth mask that contains it. Restoring a superset should subsume restoring a subset, so on those rows "necessity" cannot be read as pure evidence removal — the cited region is not always inside the mask (the models cite large boxes), and paste-back over a region that extends beyond the edit also restores authentic context the model was using. We count these rows per arm and per split rather than leave them to be discovered, and we do not claim evidence-removal semantics on them.

Evaluated under an operator the loss never saw. L_{cf} is defined with paste-back, and held-out necessity was measured with paste-back, so "faithfulness is trainable" was not yet distinguishable from "the training operator is optimisable". The inpainter audit exists on **one** split only (the permuted `seed0` split). Table 28 prints the split count per cell and omits the across-split standard deviation for single-split measurements. Under the Laplace fill, `faithalign` reaches -0.109 ± 0.000 against `detect_cite`'s -0.031 ± 0.000; under the noise-matched fill, -0.091 ± 0.000 against -0.031 ± 0.000. Section G is the reason to read those two columns with care rather than as a clean generalisation test: the same operators are sign-inverted on a detector keyed on noise energy, and one of them leaks outside the region it is given. The honest summary is that the ordering of the arms is preserved under an operator the loss never saw, on operators whose own error we have separately measured.

An honest caveat on power. 80 training and 55 held-out triples, 150 LoRA steps on a 2B model, 2 splits; the training loss for the `detect` arm oscillates late in the run, visible in the released history. With a handful of splits, the across-split standard deviations quoted above are themselves estimated from very few numbers and should be read as a spread, not as a standard error. The necessity contrasts — at the model's own citation and at the fixed ground-truth mask — are the only ones we would defend at this scale, and the second is the one we lead with.

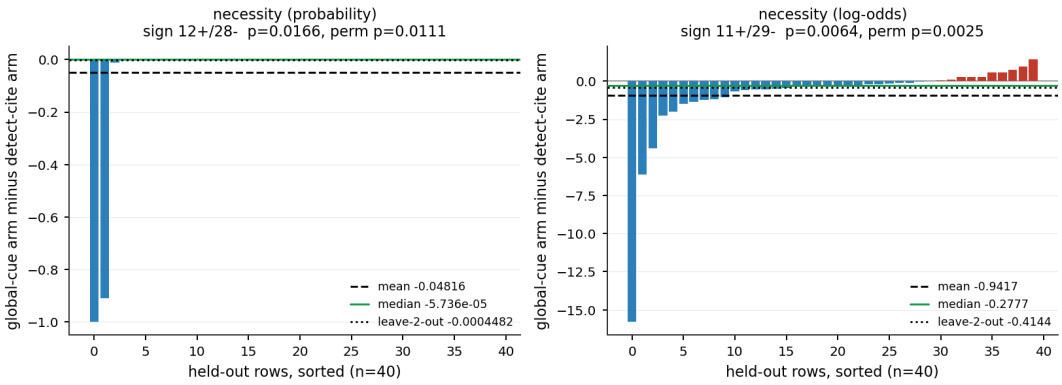


Figure 10: Edit-ID holdout; source independence unestablished. Every paired difference behind the trained global-cue contrast.

Table 26: Edit-ID holdout; source independence unestablished. The white-box paired-contrast family with Benjamini–Hochberg q-values, paired medians, exact sign tests and leave-two-out means.

Run	Contrast	Quantity	n	mean	median	sign p	perm p	q (BH)	rejects	leave-2-out
magicbrush globalcue	detect cite globalcue-detect cite	necessity	40	−0.0482	−0.00006	0.0166	0.0111	0.0406	yes	−0.00045
magicbrush globalcue	detect cite globalcue-detect cite	necessity logit	40	−0.9417	−0.27765	0.0064	0.0025	0.0125	yes	−0.41436
magicbrush globalcue	detect cite globalcue-detect cite	necessity ceiling gt	40	−0.0107	−0.00012	0.0002	0.0779	0.1947	no	−0.00028
magicbrush globalcue	detect cite globalcue-detect cite	necessity ceiling gt logit	40	−0.8727	−0.61538	0.0007	0.0001	0.0009	yes	−0.71095
magicbrush globalcue	detect cite globalcue-detect cite	localization iou	40	−0.0150	+0.00000	0.2005	0.6471	0.7810	no	−0.01101
magicbrush globalcue	detect cite globalcue-detect cite	dns matched any	39	+0.0013	+0.00001	1.0000	0.9513	0.9846	no	+0.00136
magicbrush globalcue	detect cite globalcue-detect cite	cf score matched any	39	−0.0133	+0.00000	0.3771	0.4593	0.6516	no	+0.00385
magicbrush globalcue	faithalign hinge-detect cite	necessity	40	−0.0188	+0.00003	0.6358	0.7182	0.8379	no	−0.00044
magicbrush globalcue	faithalign hinge-detect cite	necessity logit	40	−0.2692	+0.01951	1.0000	0.3796	0.6326	no	+0.03858
magicbrush globalcue	faithalign hinge-detect cite	necessity ceiling gt	40	+0.0613	+0.00775	0.0000	0.1028	0.2249	no	+0.01917
magicbrush globalcue	faithalign hinge-detect cite	necessity ceiling gt logit	40	+1.1003	+1.13362	0.0000	0.0001	0.0009	yes	+1.19443
magicbrush globalcue	faithalign hinge-detect cite	localization iou	40	−0.0348	−0.00038	0.3105	0.3368	0.5893	no	−0.00152
magicbrush globalcue	faithalign hinge-detect cite	dns matched any	39	+0.1278	+0.00054	0.1081	0.0042	0.0184	yes	+0.08038
magicbrush globalcue	faithalign hinge-detect cite	cf score matched any	39	−0.0112	−0.00016	0.3105	0.5499	0.7129	no	+0.00619
hybrid g1.0 seed0	base-detect cite localonly	necessity	40	−0.0311	−0.00049	1.0000	0.4682	0.6516	no	−0.02586
hybrid g1.0 seed0	base-detect cite localonly	necessity logit	40	−0.0048	+0.07657	0.8746	0.9846	0.9846	no	+0.25703
hybrid g1.0 seed0	base-detect cite localonly	necessity ceiling gt	40	−0.3417	−0.31193	0.0000	0.0001	0.0009	yes	−0.31333
hybrid g1.0 seed0	base-detect cite localonly	necessity ceiling gt logit	40	−1.8446	−1.54019	0.0000	0.0001	0.0009	yes	−1.63486
hybrid g1.0 seed0	base-detect cite localonly	localization iou	40	−0.0013	+0.00000	0.8679	0.9726	0.9846	no	−0.00203
hybrid g1.0 seed0	base-detect cite localonly	dns matched any	40	−0.0682	−0.01912	0.6358	0.1506	0.2842	no	−0.03215
hybrid g1.0 seed0	base-detect cite localonly	cf score matched any	40	−0.0117	+0.00000	0.8679	0.6184	0.7730	no	+0.00922
hybrid g1.0 seed0	detect cite-detect cite localonly	necessity	40	−0.0319	−0.00111	0.8746	0.4582	0.6516	no	−0.02670
hybrid g1.0 seed0	detect cite-detect cite localonly	necessity logit	40	−0.0310	+0.01056	0.8746	0.9087	0.9846	no	+0.22556
hybrid g1.0 seed0	detect cite-detect cite localonly	necessity ceiling gt	40	−0.2763	−0.19576	0.0022	0.0003	0.0017	yes	−0.24395
hybrid g1.0 seed0	detect cite-detect cite localonly	necessity ceiling gt logit	40	−0.9601	−1.30804	0.0166	0.1279	0.2633	no	−1.55712
hybrid g1.0 seed0	detect cite-detect cite localonly	localization iou	40	+0.1227	+0.02285	0.1755	0.0116	0.0406	yes	+0.08874
hybrid g1.0 seed0	detect cite-detect cite localonly	dns matched any	40	−0.0678	−0.02026	0.6358	0.1543	0.2842	no	−0.03151
hybrid g1.0 seed0	detect cite-detect cite localonly	cf score matched any	40	+0.0499	+0.00537	0.4996	0.0447	0.1304	no	+0.03204
hybrid g1.0 seed0	detect cite globalcue-detect cite localonly	necessity	40	−0.0048	−0.00025	1.0000	0.9265	0.9846	no	−0.00857
hybrid g1.0 seed0	detect cite globalcue-detect cite localonly	necessity logit	40	+0.4041	+0.10276	0.6358	0.4841	0.6516	no	+0.14902
hybrid g1.0 seed0	detect cite globalcue-detect cite localonly	necessity ceiling gt	40	−0.2764	−0.19539	0.0064	0.0003	0.0017	yes	−0.24409
hybrid g1.0 seed0	detect cite globalcue-detect cite localonly	necessity ceiling gt logit	40	−1.0233	−1.38070	0.0064	0.0923	0.2153	no	−1.59665
hybrid g1.0 seed0	detect cite globalcue-detect cite localonly	localization iou	40	+0.1128	+0.05123	0.0410	0.0151	0.0480	yes	+0.08290
hybrid g1.0 seed0	detect cite globalcue-detect cite localonly	dns matched any	40	−0.0405	−0.02032	0.8746	0.4549	0.6516	no	−0.04744
hybrid g1.0 seed0	detect cite globalcue-detect cite localonly	cf score matched any	40	+0.0469	+0.01835	0.3105	0.0488	0.1314	no	+0.03132

Two further limits belong here. First, the coverage-complete numbers come from *one* split (`permuted_seed0`, $n = 55$, the whole turn-1 corpus minus the 80 training triples): the adapters were re-audited under the new control rather than retrained, so the weights are byte-identical to the ones the split-0 run produced, but a second split under the new measurement has not been run. Second, the fragmented control is not free of bias — pooled over these arms it raises specificity by 0.0405 $[0.0074, 0.0847]$ ($p = 0.0168$, $n = 88$) on the rows where a congruent control also exists. That is roughly a seventh of the 0.2837[0.2005, 0.3849] specificity it reports for `faithalign`, so it cannot manufacture that number, but it is not zero and the conclusion should be read with it in view.

Reading Table 27. All arms: same checkpoint, same 80 training triples, same 150 steps, same seed; evaluated on a disjoint held-out split of 30 triples with the identical CF-Score ledger used for the served models. **Read the "no ctrl" column before the ΔNS and CF columns**: these models cite very large regions, so no congruent disjoint control exists on most rows and ΔNS/CF are undefined there. This table reports the *congruent-only* sensitivity analysis; the primary coverage-complete audit of the same adapters on all 55 held-out triples is in the text and in Appendix L.2. Necessity needs no control and is defined on every row.

Table 27: Edit-ID holdout; source independence unestablished. **WHITE-BOX, REAL VLM.** FaithAlign ablation on Qwen2-VL-2B (LoRA r=8 on `q_proj`/`v_proj`, 1.09M trainable parameters, 0.05% of the model).

Arm	n	AUC	p(x)-p(src)	box rate	IoU	cited area	Nec (n=30)	no ctrl	dNS	CF
base	30	0.732	0.031	0.93	0.140 [0.093, 0.196]	0.220	−0.009 [−0.017, −0.002]	15/30	−0.007	0.0546 [0.0235, 0.1006]
detect	30	1.000	0.844	0.87	0.173 [0.125, 0.233]	0.271	0.000 [0.000, 0.000]	22/30	0.000	0.0681 [0.0235, 0.1403]
detect_cite	30	1.000	0.857	1.00	0.224 [0.159, 0.305]	0.308	0.097 [0.031, 0.254]	18/30	0.240	0.1237 [0.0613, 0.2085]
faithalign	30	0.990	0.815	0.93	0.168 [0.119, 0.245]	0.358	0.255 [0.149, 0.399]	23/30	0.685	0.0759 [0.0048, 0.2379]

Table 28: Edit-ID holdout; source independence unestablished. **WHITE-BOX REPLICATION ACROSS PERMUTED SPLITS.**

Arm	splits	AUC	IoU	Nec at R(x)	Nec at M* (region-matched)	dNS	CF	Nec inpaint_diffusion	Nec inpaint_noisematched	Nec pasteback
base	2	$0.746 \pm 0.020^{[2]}$	$0.131 \pm 0.012^{[2]}$	$-0.007 \pm 0.002^{[2]}$	$0.001 \pm 0.006^{[2]}$	$-0.004 \pm 0.006^{[2]}$	$0.0460 \pm 0.0121^{[2]}$	$-0.030^{[1]}$	$-0.029^{[1]}$	$-0.007 \pm 0.002^{[2]}$
detect	2	$0.992 \pm 0.011^{[2]}$	$0.156 \pm 0.024^{[2]}$	$0.029 \pm 0.041^{[2]}$	$0.040 \pm 0.009^{[2]}$	$0.000 \pm 0.000^{[2]}$	$0.0488 \pm 0.0273^{[2]}$	$-0.000^{[1]}$	$-0.018^{[1]}$	$0.029 \pm 0.041^{[2]}$
detect_cite	2	$0.992 \pm 0.011^{[2]}$	$0.174 \pm 0.069^{[2]}$	$0.059 \pm 0.053^{[2]}$	$0.047 \pm 0.018^{[2]}$	$0.133 \pm 0.151^{[2]}$	$0.0904 \pm 0.0471^{[2]}$	$-0.031^{[1]}$	$-0.031^{[1]}$	$0.059 \pm 0.053^{[2]}$
faithalign	2	$0.984 \pm 0.008^{[2]}$	$0.152 \pm 0.022^{[2]}$	$0.261 \pm 0.009^{[2]}$	$0.273 \pm 0.167^{[2]}$	$0.368 \pm 0.449^{[2]}$	$0.0456 \pm 0.0429^{[2]}$	$-0.109^{[1]}$	$-0.091^{[1]}$	$0.261 \pm 0.009^{[2]}$

Reading Table 28. Each cell is the mean ± across-split standard deviation over 2 genuinely different train/eval splits of Qwen2-VL-2B (the aggregator refuses to pool two runs with the same eval fingerprint). **"Nec at M^*" is the region-matched column**: it restores the whole ground-truth mask, which is the identical region, operator and image set for every arm, so it cannot be inflated by an arm choosing a larger or better-placed citation. The last columns evaluate held-out necessity under the two classical inpainters, which L_{cf} never saw. **The superscript on every cell is the number of splits that cell is computed from**, and the "±" is suppressed where that number is one: the inpainter columns exist on a single split, and printing them with the row's split count and a ± 0.000 would read as agreement between two splits rather than as one measurement.

Reading Figure 11. Right: necessity at the fixed ground-truth mask — the same region for every arm, so the comparison is region- and area-matched by construction. Dots are individual splits; with a handful of splits the spread is the result, so we draw it.

Reading Figure 12. Held-out CF-Score, ΔNS and localisation IoU for the untrained checkpoint and the three nested training arms. Error bars are 95% bootstrap CIs over 30 held-out triples.

H.4 Activation patching: does the narration move with the verdict?

Having a trained model as well as an untrained one lets us ask the mechanistic question twice, before and after, on the same images.

Qwen2-VL patches the image at 14×14 and merges 2×2, so a resampled image becomes an 18×18 lattice of 324 visual tokens laid out row-major between `<|vision_start|>` and `<|vision_end|>`. That lattice is what makes "patch the cited region" well defined. We run the paste-back counterfactual $A_R(x)$ to obtain donor hidden states, then re-run the *original* image while overwriting, at one layer, exactly the visual-token activations covering the cited region R. From that single intervened forward pass we read two things: the change in $p(\text{MANIPULATED})$, and the change in the mean teacher-forced log-probability of the model's *own* explanation of the original image.

Three controls are reported with every number: the same intervention on a disjoint control region; the whole-image patch, an upper bound on what any regional patch can do; and the pixel-space counterfactual $p(x) - p(A_R(x))$, so the patch's fidelity to the intervention it was derived from is measured rather than assumed.

Control coverage and selection. A congruent-only control retains 26, 11 and 15 of 55 rows. To avoid selecting only citations with room for a translated control, the primary audit uses an area-exact fragmented fallback and records `control_kind` per row (165 rows versus 52 under the congruent-only rule; 74 congruent and 91 fragmented controls).

Congruent-only selection favours small citations. A congruent disjoint translate exists only when the citation is small enough to leave room for one, so the rule selects on citation *area*: at layer 4 the kept rows average 0.076, 0.189 and 0.252 of canvas against 0.383, 0.379 and 0.382 for the rows it discarded. On the FaithAlign arm it also selects on the *outcome*: mean pixel-space necessity is

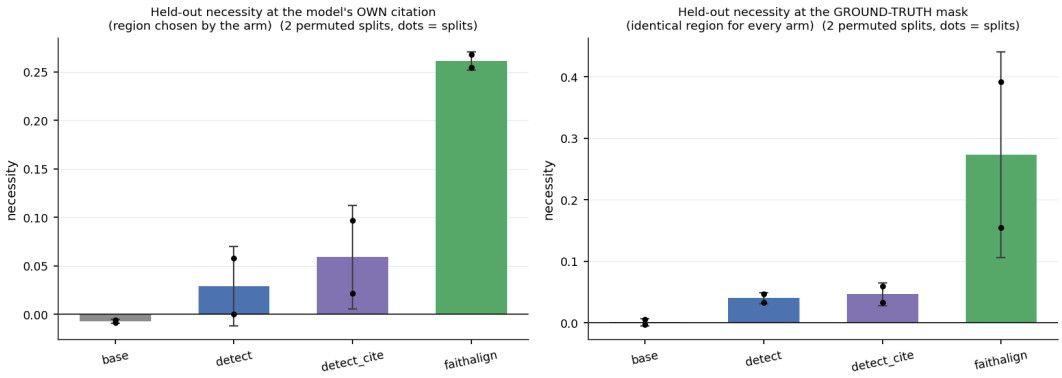


Figure 11: Edit-ID holdout; source independence unestablished. **The constructive result, with the split as the unit of replication.** Left: necessity at the region each arm chose for itself.

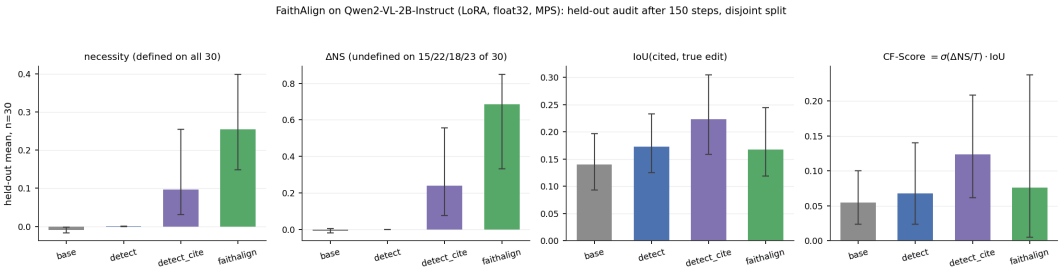


Figure 12: Edit-ID holdout; source independence unestablished. **What each loss term buys.**

$+0.544$ on the kept rows and $+0.132$ on the dropped ones, so the congruent-only subset is the subset where the pixel intervention already worked.

Whole-image activation effects increase with training. Patching at layer 4, the untrained checkpoint is inert: replacing *every* visual token's activations with the paste-back counterfactual's moves the verdict by -0.002, and the cited region alone by -0.0026 $[-0.0055, -0.0006]$ ($n = 55$). Nothing the audit can do to the image reaches the verdict, which is what an always-fake responder looks like from the inside. After `detect_cite` the whole-image patch moves it by 0.048 and after FaithAlign by 0.180: the verdict becomes a function of the image, monotonically in how much counterfactual supervision the arm received. This ordering holds under both control-coverage rules.

Control-region patches exceed cited-region patches after FaithAlign. On the biased subset the cited-region patch moved the verdict by 0.0419 $[0.0187, 0.0786]$ against 0.0113 for the control, a paired difference of 0.0306 $[0.0127, 0.0691]$ whose interval excludes zero under that selection rule. On all 55 rows the same contrast is $\mathbf{-0.0282}$ $[-0.0671, -0.0057]$ — an interval that still excludes zero and has *the opposite sign*. Patching a disjoint control moves this model's verdict *more* than patching the region it cites (0.0392 against 0.0110 $[0.0036, 0.0241]$). The stratification says where the reversal comes from and it is not noise: on the 19 congruent rows the positive cited-minus-control pattern holds (cited $+0.0343$ against control $+0.0057$), and on the 36 rows requiring fragmented controls it inverts (cited -0.0013 against control $+0.0569$). *The positive contrast is restricted to the congruent-control subset.* The same reversal appears on `detect_cite` (-0.0183 $[-0.0945, -0.0001]$, from 0.0001 $[-0.0000, 0.0003]$ on its subset), where the cited leg is null in both strata ($+0.0001$ and $+0.0001$) and the control leg is not (-0.0000 against $+0.0338$).

Interpretation. What the counterfactual objective demonstrably does is raise the verdict's dependence on the visual tokens *as a whole* (the $-0.002 \rightarrow 0.180$ ladder above). What we cannot show is that it puts that dependence on the region the model cites: on the rows where the citation is large — which are two thirds of them, and the ones an audit most needs — a disjoint region carries more of the verdict than the citation does. This is a negative result for regional alignment under the evaluated FaithAlign objective.

Narration–verdict coupling remains unresolved. The same intervened forward passes say almost nothing changed for the explanation. The drop in the model's own explanation log-probability is 0.0051 $[0.0007, 0.0216]$ before training and 0.0094 $[0.0053, 0.0162]$ after, with a contrast whose interval includes zero, while the whole-image verdict effect grew by two orders of magnitude over the same interval. Per example the verdict shift and the explanation shift are essentially uncorrelated at every depth and in every arm: Spearman $\rho = -0.09$ (base), -0.01 (`detect_cite`) and 0.01 (`faithalign`). This leg does not condition on the control's existence. L_{cf} constrains p and nothing in it constrains the decoder, so the narration is still produced by a pathway the verdict's evidence does not reach; the obvious next term, a counterfactual constraint on the explanation's likelihood rather than on the verdict alone, we have not run.

Sample size and uncertainty. n is 55 on every arm, all 55 attempted, with which control was used recorded per row. The coupling correlations still have wide intervals and are reported as exploratory. The two contrasts whose intervals exclude zero are both *negative* cited-minus-control differences (-0.0282 $[-0.0671, -0.0057]$ and -0.0183 $[-0.0945, -0.0001]$), which is the reverse of the direction the congruent-only control gives and is the finding of this subsection.

Reading Table 29. ΔVerdict is $p(x)$ minus p under the patch; ΔExplain is the drop in the mean log-probability of the model's own explanation under the *same* intervened forward pass. $\rho(V, E)$ is the Spearman correlation between them across examples — the coupling of the two readouts.

Reading Figure 13. Left: the effect on the verdict of replacing visual-token activations with the paste-back counterfactual's, at three depths, for the cited region, a disjoint control region (congruent where one exists, area-exact fragmented otherwise) and the whole image. **The control bar exceeds the cited bar**, unlike the contrast restricted to congruent-control rows. Right: per-example verdict shift against explanation-likelihood shift under the same intervention.

Reading Figure 14. Note the y-axis scale against Figure 13.

I Separating the pathology from the corpus

Every number in the two preceding sections comes from one corpus produced by one generator. That limits interpretation because a DALL-E 2 edit re-encodes the whole canvas, so a MagicBrush manipulated image differs from its authentic source *everywhere*, with median outside-mask mean absolute difference $0.9988/255$ and a minimum of $0.0123/255$ over 262 triples — never zero. We then built the global-cue oracle out of exactly that fingerprint. A canvas-wide *sufficient* cue mechanically implies that an accurate detector *can* have near-zero *local* necessity. So the headline empirical finding — accurate detectors whose verdicts do not depend on the region they cite — cannot, on MagicBrush alone, be separated from the alternative explanation "on this corpus the evidence is not local". That distinction is the whole point of the paper, and no experiment above tests it.

A corpus where the evidence is local by construction. We build one from the same pixels. Taking only the *authentic* MS-COCO photographs of MagicBrush (no DALL-E-2 pixel is ever read) and the real ground-truth region shapes, we compose two classical manipulation families: **splice** (the region is filled with content from a different photograph) and **copy-move** (the region is filled from the same photograph at a translated offset). The composite is done in `uint8` and stored as lossless PNG, so

$$\left| x_{\mathrm{tgt}}(u) - x_{\mathrm{src}}(u) \right| = 0 \quad \text{exactly, for every } u \notin M^{\star}. \tag{19}$$

This is verified on every stored row after writing, from the PNGs rather than from memory: the maximum outside-mask residual over the whole corpus is $0.0/255$ on 180 triples (90 splice, 90 copy-move), median inside-mask MAD $71.2/255$ (Fig. 16).

Disclosure: how the paste is made, and what that costs. The initial corpus is a **hard uint8 paste with no feathering, no colour or noise matching between donor and host, and no recompression** — the manifest records exactly that string. It matters, because the corpus was built to remove one confound (the canvas-wide generator fingerprint) and silently introduced another: a hard discontinuity at the region boundary is a trivially *local* cue, which is consistent with `qwen2.5vl:7b` reaching AUC 0.992 here against 0.868 on MagicBrush. We therefore build and ship two further variants from the same script, and disclose the measured cost of each:

Table 29: Edit-ID holdout; source independence unestablished. **WHITE-BOX, REAL VLM.** Visual-token activation patching on Qwen2-VL-2B, before and after training, on held-out MagicBrush triples ($n = 26$, 11 and 15 of 55 attempted; rows without a congruent disjoint control are dropped).

Arm	layer	dVerdict cited	dVerdict control	dVerdict whole	dExplain cited	rho(V,E)
base	4	−0.0026 [−0.0055, −0.0006]	−0.0024	−0.0020	0.0051	−0.085
base	14	−0.0007 [−0.0012, −0.0003]	−0.0015	−0.0036	0.0038	−0.025
base	24	−0.0001 [−0.0004, 0.0000]	−0.0006	−0.0004	0.0010	0.009
detect_cite	4	0.0001 [0.0000, 0.0002]	0.0184	0.0481	0.0137	−0.006
detect_cite	14	−0.0000 [−0.0000, 0.0000]	−0.0000	0.0000	0.0104	−0.110
detect_cite	24	−0.0000 [−0.0000, 0.0000]	−0.0000	0.0000	0.0012	−0.008
faithalign	4	0.0110 [0.0036, 0.0241]	0.0392	0.1795	0.0094	0.011
faithalign	14	−0.0004 [−0.0010, 0.0001]	0.0007	−0.0012	0.0062	−0.065
faithalign	24	−0.0002 [−0.0008, 0.0001]	−0.0001	−0.0001	0.0017	0.023

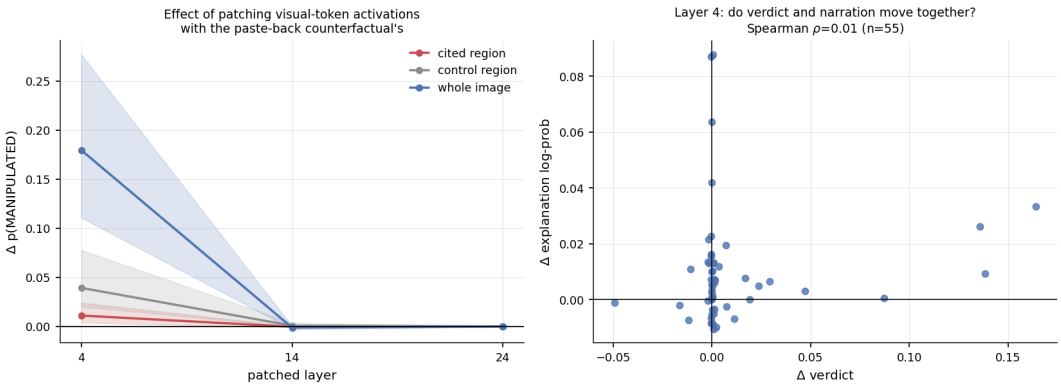


Figure 13: Edit-ID holdout; source independence unestablished. **Activation patching on a real VLM, after FaithAlign ($n = 55$, coverage-complete controls).**

- **Feathered.** An alpha ramp of 3 pixels kept *strictly inside* $M^\star$ (alpha is exactly 0 outside the mask, so the product with the host is exact for integers). Eq. 19 therefore still holds with equality — the maximum outside-mask residual is 0 on every stored row — while the seam is no longer a step. This is the variant the hybrid corpus (§J) is built on.
- **JPEG-75.** Both stored images re-encoded at quality 75, which *breaks* bit-identity on purpose so the residual can be measured instead of asserted. Measured on the stored files: outside-mask deviation median 0.102 levels, maximum 68.7 levels (the maximum sits in the 8×8 blocks straddling the region boundary, which is where a real recompression leaks). Eq. 19 becomes an approximation there, and $\mathrm{Nec}(M^\star) = p(x) - p(x_{\mathrm{src}})$ becomes an identity up to that residual rather than exactly.

Does the result survive the construction? We re-ran the identical served audit — same model, same 31-call ledger, same 50 host photographs — on the variants (Table 30): all three construction variants were audited. The number at issue is the cited-region/full-region normalized intervention ratio, 0.335 $[0.216, 0.472]$ on the hard-paste corpus. It is not an additive allocation of internal evidence. Feathering (which preserves exact locality) makes it *not estimable* by our own quoting rule (ratio-of-means 0.205 against mean-of-ratios -0.321 $[-1.889, 0.137]$ — opposite signs, so neither is reported): AUC 0.877 against 0.992, $\mathrm{Nec}(M^\star)$ 0.5181 $[0.4158, 0.6187]$ against 0.7250 $[0.6337, 0.7938]$, measured identity error 0.00e+00. Paired against the hard-paste corpus on the same 50 host photographs: $\mathrm{Nec}(M^\star)$ -0.2068 $[-0.2995, -0.1241]$, $\mathrm{Nec}(R(x))$ -0.1305 $[-0.2563, -0.0101]$. JPEG-75 recompression (which breaks it) moves it to 0.449 $[0.252, 0.853]$: AUC 0.976 against 0.992, $\mathrm{Nec}(M^\star)$ 0.5810 $[0.4773, 0.6742]$ against 0.7250 $[0.6337, 0.7938]$, measured identity error 8.56e-01. Paired against the hard-paste corpus on the same 50 host photographs: $\mathrm{Nec}(M^\star)$ -0.1440 $[-0.2171, -0.0743]$, $\mathrm{Nec}(R(x))$ 0.0036 $[-0.0777, 0.0807]$.

Construction sensitivity. Three findings qualify the result.

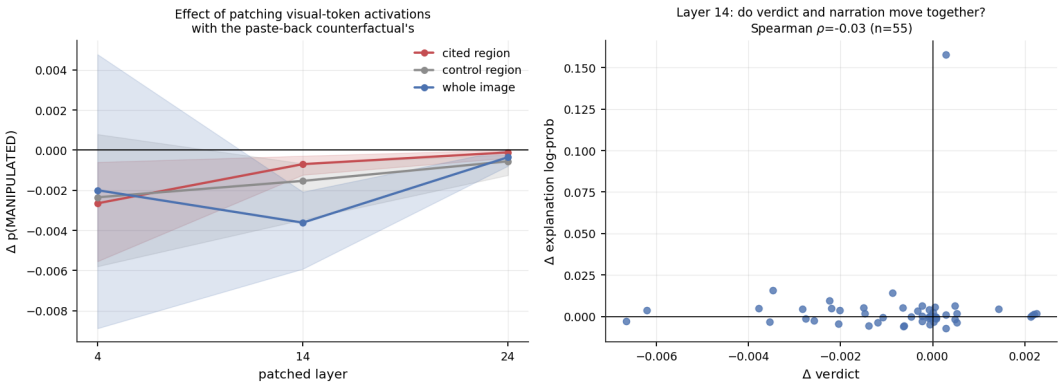


Figure 14: Edit-ID holdout; source independence unestablished. **The same intervention on the untrained checkpoint.** Nothing moves: replacing every visual token with the counterfactual's activations shifts the verdict by -0.002.

(a) The seam was carrying a lot. Blending it away — while keeping locality *exact*, which the identity-error column verifies — drops detection from 0.992 to 0.877 and the necessity ceiling from 0.7250 $[0.6337, 0.7938]$ to 0.5181 $[0.4158, 0.6187]$. A meaningful part of both the detectability and the locally recoverable evidence in the initial corpus was a hard discontinuity.

(b) The headline ratio is not estimable on the blended corpus. Its two estimators take opposite signs there, so by the rule this paper set for itself in the prespecified protocol we report neither as a stable normalized effect. The hard-paste estimate 0.335 $[0.216, 0.472]$ is construction-specific, not a general model property.

(c) It does survive recompression. On the JPEG-75 variant the two estimators agree and the ratio is 0.449 $[0.252, 0.853]$, overlapping the hard-paste 0.335 $[0.216, 0.472]$; necessity at the cited region is unchanged (0.0036 $[-0.0777, 0.0807]$ paired), and detection barely moves (0.976). The price is exactly the one Eq. 19 predicts: the identity $\mathrm{Nec}(M^\star) = p(x) - p(x_{\mathrm{src}})$ stops being exact, with a measured maximum error of 8.56e-01 — **about the same size as the error DALL-E 2's canvas re-encode induces on MagicBrush** (0.8560) — and the fraction of rows on which specificity is *exactly* zero falls from 1.00 to 0.48. A JPEG-75 round trip costs about as much of the exactness as a generative re-encode does, which is a useful calibration for anyone applying paste-back to a corpus that has been through a codec.

The directional comparison holds on all three constructions: restoring the true mask recovers a large part of the verdict while restoring the region the model *cites* recovers much less ($\mathrm{Nec}(R(x))$ 0.2369 $[0.1486, 0.3491]$ against $\mathrm{Nec}(M^\star)$ 0.7250 $[0.6337, 0.7938]$ hard-paste; 0.1063 $[0.0123, 0.2115]$ against 0.5181 $[0.4158, 0.6187]$ feathered; 0.2405 $[0.1459, 0.3603]$ against 0.5810 $[0.4773, 0.6742]$ under JPEG).

Reading Table 30. "recovery" is the cited-region/full-region normalized intervention ratio — not an evidence share — and the last two columns are the corpus properties the construction is supposed to guarantee, measured on the stored files rather than asserted.

Eq. 19 has two consequences we can check rather than argue about.

(i) Paste-back is exact, so the necessity ceiling *is* the detection margin. $A_{M^\star}(x)$ is bit-for-bit x_{src}, hence $\mathrm{Nec}(M^\star) = p(x) - p(x_{\mathrm{src}})$ identically. A detector on this corpus *cannot* have a large detection margin and a small ceiling necessity; the whole verdict separation is, by construction, recoverable by a local intervention. We verify the identity numerically rather than assert it: the largest absolute discrepancy over all served rows is $0 exactly$, against 0.8560 on MagicBrush — that second number *is* the fingerprint, expressed as the part of the verdict that no local intervention can reach.

(ii) A detector that reads only outside the citation is blind. The global-cue oracle's statistic is the residual against the source outside the cited box; on this corpus that residual is identically zero for authentic and manipulated images alike. Its AUC must therefore fall from 1.000 to chance,

Table 30: The same audit, the same model and the same host photographs on three constructions of the local-evidence corpus.

Corpus variant	n	AUC	IoU	Nec at R(x)	Nec at M*	recovery (RoM)	|Nec(M*) - margin| max	frac Spec exactly 0
hard paste (initial corpus)	50	0.992	0.353 [0.263, 0.450]	0.2369 [0.1486, 0.3491]	0.7250 [0.6337, 0.7938]	0.327	0.00e+00	1.00
feathered (3 px, inside M*)	50	0.877	0.234 [0.157, 0.331]	0.1063 [0.0123, 0.2115]	0.5181 [0.4158, 0.6187]	0.205	0.00e+00	1.00
JPEG-75 re-encode	50	0.976	0.378 [0.289, 0.472]	0.2405 [0.1459, 0.3603]	0.5810 [0.4773, 0.6742]	0.414	8.56e-01	0.48

and it does: 1.000 on MagicBrush against 0.500 here. The same holds for the *training* counterpart introduced in Section H: the global-cue LoRA arm is not merely weak on this corpus, it is *undefined* — its two training classes are the same image — and the trainer raises rather than silently degrades.

The positive control replicates on the new corpus, and the global-cue oracle stops being a detector. We re-ran the oracle families on the local corpus (60 triples $\times$ 3 seeds, 1800 rows, CPU only, zero network). The metric validation reproduces: faithful 0.753[0.734, 0.769] against unfaithful 0.000[0.000, 0.001], a paired gap of 0.7534[0.7358, 0.7702] ($p < 10^{-4}$), monotone in the designed citation faithfulness (pooled Spearman $\rho = 0.978$). This matters independently of the corpus argument: it is the first replication of the positive control on a corpus that shares no manipulated pixel with the one it was designed on. The faithful arm's necessity is 0.9928 here against 0.9914 there.

The global-cue arm behaves exactly as Eq. 19 says it must: its AUC falls from 1.000 to 0.500 — chance — because the statistic it reads is identically zero on both classes. One thing worth stating plainly, because it is a property of the metric rather than of the corpus: CF-Score still assigns that arm 0.384[0.374, 0.392], since its IoU is maximal and $\Delta\mathrm{NS} \approx 0$ pins the sigmoid at $\frac{1}{2}$. CF-Score is a faithfulness metric conditional on detection and does not penalise a non-detector; the detection stratification of Section L.2 is what handles that, and this arm is a clean demonstration of why that stratification is not optional.

The sharpest corpus signature is a count of exact zeros, not an interval. On a corpus whose evidence is local, restoring a region *disjoint* from $M^\star$ rewrites no pixel at all — paste-back is literally the identity there — so specificity is exactly 0. On a corpus with a canvas-wide fingerprint, restoring *any* region removes some of that fingerprint, so specificity is never exactly 0. That is what we measure on `qwen2.5vl:7b`: the congruent-control specificity is exactly zero on 70% of the 37 local-corpus rows that have one, against 0% of 39 on MagicBrush — not "small on MagicBrush", *never zero on any row*. At the ceiling's own control region, which is disjoint from $M^\star$ by construction, specificity is exactly zero on 100% of 50 local rows, so on this corpus $\Delta\mathrm{NS}(M^\star) = \mathrm{Nec}(M^\star)$ identically and the ceiling claim needs no statistical control at all: a model that responded to paste-back *anywhere* would show up immediately as a non-zero number in a column that is otherwise exactly zero.

What the corpus contrast shows. Table 31 reports the identical design on both corpora. The finding is neither of the two clean stories, and we state it as measured rather than as either.

The pathology is partly, but only partly, corpus-induced. `qwen2.5vl:7b` detects these manipulations *better* than the DALL-E-2 ones (AUC 0.992 on 50 triples against 0.868 on 50), and its necessity ceiling rises from 0.1236 $[0.0766, 0.1836]$ to 0.7250 $[0.6337, 0.7938]$ — as it must, since here the ceiling *is* the detection margin. Necessity at the region the model actually cites also rises, from 0.0292 $[-0.0236, 0.0886]$ to 0.2369 $[0.1486, 0.3491]$. So a substantial part of the initial finding that "restoring the true mask barely moves the verdict" was the corpus: on MagicBrush most of the verdict lives in pixels no local intervention reaches, with a measured residual (0.8560 of separation unreachable at worst).

What does not go away is the gap between the ceiling and the citation. On the local corpus the model recovers 0.2369 $[0.1486, 0.3491]$ of an *available* 0.7250 $[0.6337, 0.7938]$ at the region it points at: a fraction of 0.335 $[0.216, 0.472]$ (mean of per-row ratios over all 50 rows, no row dropped; ratio of means 0.327, in agreement, which is the condition under which we quote a ratio at all). **Read it as a property of this hard-paste construction, not of the model**: the ratio is not estimable on the blended variant. Every restoration effect here is attributable to manipulated pixels and nothing else, because restoring anything else is provably a no-op. The result therefore concerns

local evidence: restoring the model's cited region recovers about a third of the effect of restoring $M^\star$ — a ratio of two interventions, not a third of the evidence.

FakeShield on the same corpus. Section E.3 evaluates FakeShield v1-22B (DTE-FDM) through the identical instrument. There its ceiling is 0.2856 $[0.2468, 0.3247]$, an identity rather than an estimate, and the citation effect whose interval included zero on MagicBrush becomes 0.0233 $[0.0050, 0.0563]$, excluding zero, at 8% $[1\%, 17\%]$ of the ceiling — against 0.335 $[0.216, 0.472]$ for the chat model on the same corpus. This comparison shows why the MagicBrush citation result cannot be generalised across corpora.

And the same ratio is not even estimable on MagicBrush — which is the corpus effect, expressed as a statistic that refuses to be computed. 16 of 50 MagicBrush rows have a ceiling below 0.02: on those images there is essentially nothing local to recover, so the denominator is noise. The mean of per-row ratios is -0.716 $[-2.835, 0.236]$ against a ratio of means of 0.236 — they disagree in sign, so by our own rule we do not quote either. On the local corpus no row is dropped and the two estimators agree to 0.008. A reader who wants one sentence for the whole section can have this one: *the corpus determined how much of the verdict was locally recoverable, and the model determined how much of that it pointed at; only the first was ever a property of DALL-E 2.*

Reading Table 31. "Nec at M^*" is the necessity ceiling; on the local corpus it equals the detection margin by construction. "dNS at M^*" subtracts from this necessity the effect at an area-exact control region placed elsewhere and restored with the same operator — without it, "restoring M^* moves the verdict" cannot be told apart from "restoring anything of that size moves the verdict".

The white-box pathology is *not* corpus-induced, and on this corpus it is far starker. We reran the whole FaithAlign ablation on the local corpus: same checkpoint (Qwen2-VL-2B), same 150 steps, same seed, same audit ledger, 60 training and 40 evaluation triples, disjoint by edit ID; source/donor independence is unestablished. The `detect` arm — trained on nothing but the verdict — reaches AUC 0.951 and has a held-out necessity at the *ground-truth mask* of 0.8277 $[0.6910, 0.9224]$. That number is forced: on this corpus the ceiling is the detection margin. What is not forced is what happens at the region the model *cites*, where its necessity is 0.0756 $[-0.0232, 0.2003]$.

That experiment comes out against the corpus explanation. On MagicBrush a `detect` arm's held-out necessity of 0.029 ± 0.041 was fully compatible with "there is little local evidence to find". Here there is a great deal of local evidence — 0.8277 $[0.6910, 0.9224]$ of it, recoverable by a single paste-back, on images where restoring anything else is a provable no-op — and the model still points somewhere that recovers a small fraction of it. Untrained, the same checkpoint reaches AUC 0.895 with ceiling necessity 0.0840 $[0.0650, 0.1112]$; training the verdict raises the ceiling by an order of magnitude and leaves the citation where it was.

Two results about *training* that only this corpus could produce, one positive and one a clean null. (a) Explanation imitation buys real faithfulness here, and did not there. Adding L_{cite} moves localisation from 0.2341 $[0.1680, 0.3107]$ to 0.4202 $[0.3257, 0.5185]$ and necessity at the model's own citation from 0.0756 $[-0.0232, 0.2003]$ to 0.2768 $[0.1617, 0.4223]$ — a paired gain of $+0.2011$ $[+0.0448, +0.3652]$ ($p = 0.014$, $n = 40$), with CF-Score up $+0.1691$ $[+0.0818, +0.2627]$ ($p = 0.0003$). On MagicBrush the same contrast was small and not separable from noise. The reason is visible in the ceiling: there, teaching the model where to point could not help much, because most of the verdict lived where no citation could reach it.

(b) The additional counterfactual-training effect is unresolved. Against `detect_cite`, `faithalign` moves necessity by -0.0013 $[-0.0057, +0.0020]$ ($p = 0.54$), ceiling necessity by $+0.0009$ $[-0.0015, +0.0030]$ ($p = 0.47$), IoU by -0.0001, and CF-Score by -0.0005. Those are not underpowered intervals; they are intervals two orders of magnitude tighter than the effect the same term produced on MagicBrush ($+0.2465$[0.1603, 0.3342](p=0.0001,n=55) in necessity, $+0.3320$[0.2453, 0.4231](p=0.0001,n=55) at the ceiling). **This is consistent with the MagicBrush control-effect diagnosis.** We concluded there that L_{cf} "taught the model to respond to gold paste-back *anywhere*", which raises necessity and specificity together. On a corpus where paste-back over any region disjoint from $M^\star$ is the *identity map*, there is no "anywhere" to become sensitive to — and the term's entire measured effect vanishes. The contrast limits the training result in §H.1:

Table 31: Edit-ID holdout; source independence unestablished. **THE SAME DESIGN ON TWO CORPORA.** MagicBrush (DALL-E-2 edits, canvas-wide fingerprint) against the local-evidence corpus (outside-mask difference exactly 0).

system	corpus	n	AUC	IoU	Nec at R(x)	Nec at M*	dNS at M*
served qwen2.5vl:7b	magicbrush	50	0.868	0.206 [0.150, 0.277]	0.029 [−0.024, 0.089]	0.124 [0.077, 0.184]	0.098 [0.034, 0.163]
served qwen2.5vl:7b	splice	50	0.992	0.353 [0.263, 0.450]	0.237 [0.149, 0.349]	0.725 [0.634, 0.794]	0.725 [0.634, 0.794]
Qwen2-VL-2B base	magicbrush	55	0.760	0.123 [0.091, 0.166]	−0.006 [−0.011, −0.001]	0.005 [−0.001, 0.011]	–
Qwen2-VL-2B detect	magicbrush	55	0.984	0.139 [0.104, 0.183]	0.058 [0.020, 0.148]	0.047 [0.010, 0.131]	–
Qwen2-VL-2B detect_cite	magicbrush	55	0.984	0.125 [0.091, 0.169]	0.022 [−0.017, 0.093]	0.060 [0.006, 0.149]	–
Qwen2-VL-2B faithalign	magicbrush	55	0.978	0.137 [0.097, 0.189]	0.268 [0.192, 0.361]	0.392 [0.298, 0.494]	–
Qwen2-VL-2B base	splice	40	0.895	0.210 [0.148, 0.285]	0.008 [0.003, 0.015]	0.084 [0.065, 0.111]	0.084 [0.065, 0.111]
Qwen2-VL-2B detect	splice	40	0.951	0.234 [0.168, 0.311]	0.076 [−0.023, 0.200]	0.828 [0.691, 0.922]	0.828 [0.691, 0.922]
Qwen2-VL-2B detect_cite	splice	40	0.958	0.420 [0.326, 0.519]	0.277 [0.162, 0.422]	0.772 [0.640, 0.869]	0.772 [0.640, 0.869]
Qwen2-VL-2B faithalign	splice	40	0.958	0.420 [0.326, 0.518]	0.275 [0.160, 0.421]	0.773 [0.642, 0.870]	0.773 [0.642, 0.870]

improved response to the MagicBrush restoration does not establish improved alignment with local evidence on a different corpus.

Reading Figure 15. Solid bar: necessity at the region the system actually cites. The pale bar is a property of the corpus (on the local corpus it *is* the detection margin, by construction). The gap between the two is the property of the model, and it does not close.

Reading Figure 16. Top two rows: one splice and one copy-move triple from the local-evidence corpus; the difference panel is black outside the mask *exactly*. Third row: a MagicBrush DALL-E-2 edit, whose difference panel is non-zero over the entire canvas. Bottom: the outside-mask residual distribution for both corpora. The MagicBrush distribution is what the global-cue oracle reads; the local corpus is a spike at exactly zero on all 180 rows.

J THE HYBRID CORPUS IN FULL

J.1 THE RESOLUTION–GLOBAL-CUE TRADE-OFF

The two base corpora present a structural trade-off.

MagicBrush edits are DALL-E-2 inpaintings, and DALL-E-2 re-encodes the whole canvas: the prepared corpus measures an outside-mask deviation with median 0.9988/255 and minimum 0.0123/255 and *never* zero. That canvas-wide fingerprint is a sufficient cue for "manipulated", so a detector can be accurate while its verdict depends on nothing local — exactly the pathology this paper is about. But paste-back at the ground-truth region leaves the fingerprint untouched, so on that corpus *every* arm's necessity lives in a $\sim 10^{-4}$ band and the instrument has no resolution.

The splice/copy-move corpus (§I) fixes the resolution: paste-back is exact, so $\mathrm{Nec}(M^\star) \equiv p(x) - p(x_{\mathrm{src}})$ and necessity has the full detection margin as its range. But there is no canvas-wide cue at all, so the correct-but-non-causal arm cannot be trained: its two training classes are literally the same image, and our trainer raises rather than silently producing a degenerate arm.

So the dangerous cell could only be exhibited where the instrument was blind, and the instrument only had resolution where the cell could not exist. The hybrid design combines the two properties.

J.2 MANUFACTURING THE CUE AT A KNOWN STRENGTH

We dissolve the bind by building the cue ourselves. On top of the exact-local corpus we add a canvas-wide *generator fingerprint* to the positive class only:

$$F_\gamma(I) \;=\; \mathrm{clip}_{[0,255]}\Big(I \;+\; \gamma\, A_c\,(-1, 0, +1) \;+\; \gamma\, A_g\, \mathrm{sign}\big(\sin\tfrac{2\pi y}{P} \sin\tfrac{2\pi x}{P}\big)\Big), \tag{20}$$

with $A_c = 10$, $A_g = 6$ levels and $P = 16$ pixels. Two properties are what make it usable as an instrument rather than as decoration, and both are asserted by tests rather than argued:

- **Uniform energy.** Both terms have constant magnitude at every pixel — the sign is what buys that, since a plain sinusoid would concentrate energy at its antinodes and be weakly localisable. Measured inside-mask/outside-mask energy ratio is 1.00 to within 0.05. Restoring authentic pixels over a region of area a therefore removes approximately a fraction a of

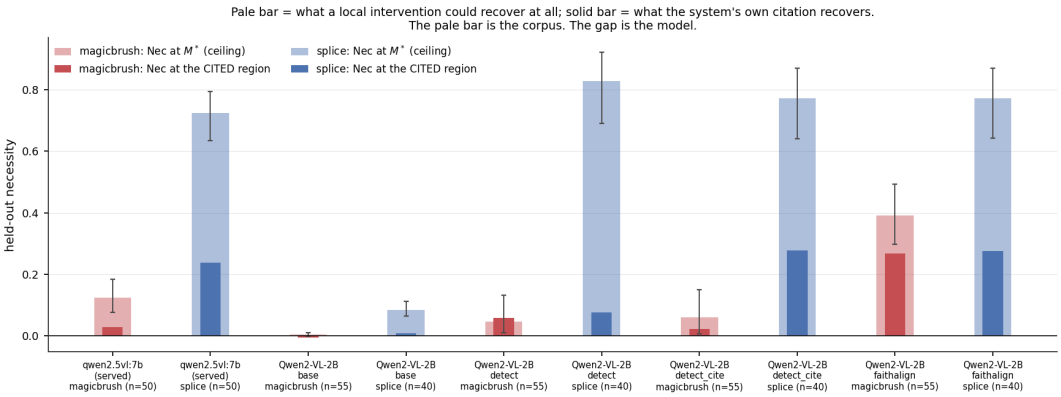


Figure 15: Edit-ID holdout; source independence unestablished. **The corpus moved the ceiling; the model kept the gap.** Pale bar: necessity at the ground-truth mask — everything a local intervention could recover.

the cue, which is the quantitative content of "a canvas-wide cue survives a local intervention".

- **Known, linear strength.** γ scales it linearly and $\gamma = 0$ is the identity *bit-for-bit*, so the sweep's zero point is the untouched corpus. Measured outside-mask deviation on the stored PNGs is 2.16, 4.31 and 8.57 levels at $\gamma = 0.25, 0.5, 1.0$ — i.e. the weakest setting is already about twice the DALL-E-2 fingerprint MagicBrush carries, and the strongest about eight times it.

The corpus stores a fourth image, $F_\gamma(x_{\text{src}})$: the authentic host carrying the same fingerprint and *no* manipulation. It is never audited. It exists so that each of the two cues can be made uninformative for one training arm:

arm	verdict supervision	cue it can use
`detect_cite_localonly`	x vs $F_\gamma(x_{\text{src}})$	local evidence only
`detect_cite_globalcue`	$F_\gamma(x_{\text{src}})$ vs x_{src}	canvas fingerprint only
`detect_cite`	x vs x_{src}	both

All three receive *identical* citation supervision (the same reference string, the same ground-truth box, the same weight), the same LoRA configuration, the same 120 steps, the same learning rate and the same edit-ID holdout split, and all three are audited on the *identical* (x, x_{src}) pair with the identical audit ledger. The intent is that they differ only in *mechanism*; whether that intent survived contact with the training dynamics is measured rather than assumed, and it did not (§J.3). What the corpus does deliver, and what MagicBrush could not, is a setting in which necessity has room to move at all.

J.3 Result: the fingerprint sweep

Table 32 gives every arm at every γ that has been run, and Table 33 every paired contrast.

Reading Table 32. The two one-cue arms share their citation supervision exactly; the IoU column is the control that was *supposed* to make the necessity columns interpretable, and §J.3 reports what happened to it.

Reading Table 33. Mean with its bootstrap interval, *paired median*, exact sign test, permutation p, the Benjamini–Hochberg q over the hybrid family, and the leave-two-out mean — so a mean carried by a small number of rows is visible rather than inferable.

Reading Figure 17. Left: whether the canvas-wide cue is learnable at all, as γ grows. Right: whether each arm's verdict depends on the manipulated region, measured at the region-matched $M^\star$. The sweep is *not* entailed by construction — nothing in the design forces the global-cue arm to reach

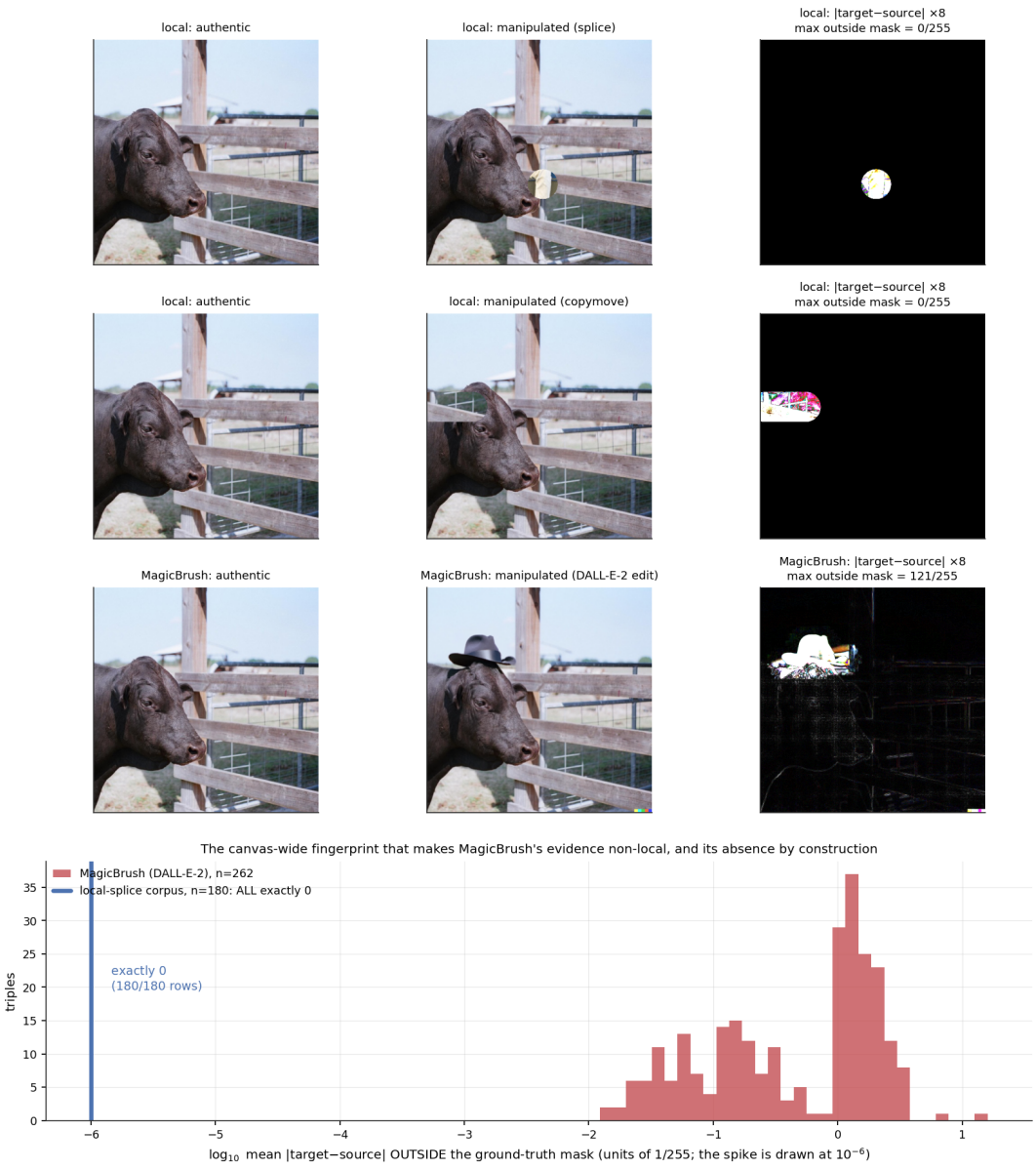


Figure 16: Edit-ID holdout; source independence unestablished. **The alternative explanation, and its removal.**

any particular AUC, and the γ at which it becomes trainable is a measurement of what a 2B VLM can extract from a canvas signature of that strength. **The figure is withheld when fewer than two settings have completed**: one point is not a curve, and plotting it would read as evidence about a shape we have not measured.

The primary contrast is region-matched, and here is why. The design intends the two arms to differ only in *mechanism*, with citation held fixed by identical supervision. On MagicBrush, the IoU contrast was unresolved; equivalence was not established. **On the hybrid corpus it does not**, and we report that before the result rather than after it. At $\gamma = 1$ the global-cue arm's IoU is 0.250 $[0.174, 0.343]$ against the local-evidence arm's 0.137 $[0.085, 0.211]$ (paired difference 0.1128 $[0.0288, 0.1996]$, $p = 0.0151$). The mechanism explains it: the global-cue arm's detection task is much easier — its verdict loss reaches near-zero within 30 steps — so more of its budget goes to the

Table 32: Edit-ID holdout; source independence unestablished. Hybrid corpus, Qwen2-VL-2B LoRA arms, $n = 40$ held-out triples per cell.

gamma	seed	Arm	n	AUC	IoU	Nec at R(x)	Nec logit	Nec at M*	p(x) mean
1.00	0	base checkpoint	40	0.999	0.136 [0.092, 0.192]	0.003 [0.002, 0.004]	0.29 [0.17, 0.42]	0.010 [0.004, 0.032]	0.9914
1.00	0	local-evidence arm	40	1.000	0.137 [0.085, 0.211]	0.034 [−0.040, 0.121]	0.29 [−0.12, 0.84]	0.351 [0.240, 0.466]	0.5201
1.00	0	global-cue arm	40	1.000	0.250 [0.174, 0.343]	0.029 [0.000, 0.147]	0.70 [0.22, 2.31]	0.075 [0.025, 0.199]	0.9994
1.00	0	both-cues arm	40	1.000	0.260 [0.180, 0.357]	0.002 [0.000, 0.011]	0.26 [0.09, 0.56]	0.075 [0.025, 0.200]	0.9997
1.00	1	local-evidence arm	40	1.000	0.186 [0.125, 0.259]	0.087 [0.048, 0.152]	0.78 [0.47, 1.17]	0.328 [0.243, 0.422]	0.9141
1.00	1	global-cue arm	40	1.000	0.267 [0.186, 0.358]	−0.000 [−0.000, 0.000]	−0.04 [−0.15, 0.07]	0.050 [0.000, 0.175]	0.9999
1.00	2	local-evidence arm	40	1.000	0.243 [0.173, 0.326]	0.120 [0.058, 0.196]	0.71 [0.38, 1.11]	0.394 [0.302, 0.485]	0.7575
1.00	2	global-cue arm	40	1.000	0.336 [0.249, 0.429]	0.007 [0.000, 0.035]	0.36 [0.10, 0.87]	0.073 [0.023, 0.194]	0.9998
0.50	0	local-evidence arm	40	0.960	0.210 [0.132, 0.305]	0.167 [0.079, 0.281]	1.35 [0.58, 2.31]	0.592 [0.440, 0.709]	0.7705
0.50	0	global-cue arm	40	0.984	0.243 [0.167, 0.333]	0.053 [0.008, 0.139]	0.63 [0.08, 1.35]	0.266 [0.186, 0.359]	0.9139
0.25	0	local-evidence arm	40	0.884	0.240 [0.156, 0.337]	0.109 [−0.005, 0.244]	1.10 [0.14, 2.35]	0.656 [0.513, 0.777]	0.7169
0.25	0	global-cue arm	40	0.838	0.214 [0.143, 0.311]	0.027 [0.010, 0.048]	0.12 [0.04, 0.21]	0.105 [0.078, 0.138]	0.6032

Table 33: Edit-ID holdout; source independence unestablished. Global-cue arm minus local-evidence arm, paired on held-out rows.

gamma	seed	Quantity	n	mean	median	sign +/-	sign p	perm p	q (BH)	leave-2-out
1.00	0	Nec	40	−0.0048 [−0.0998, 0.0881]	−0.0003	20/20	1.0000	0.9265	0.9265	−0.0086
1.00	0	Nec (logit)	40	0.404 [−0.404, 1.412]	0.1028	22/18	0.6358	0.4841	0.5233	0.1490
1.00	0	IoU	40	0.1128 [0.0288, 0.1996]	0.0512	24/11	0.0410	0.0151	0.0403	0.0829
1.00	1	Nec	40	−0.0875 [−0.1421, −0.0403]	−0.0234	12/28	0.0166	0.0006	0.0022	−0.0596
1.00	1	Nec (logit)	40	−0.819 [−1.189, −0.458]	−0.4651	14/26	0.0807	0.0001	0.0004	−0.6774
1.00	1	IoU	40	0.0807 [−0.0266, 0.1890]	0.0294	23/14	0.1877	0.1507	0.2232	0.0459
1.00	2	Nec	40	−0.1134 [−0.1877, −0.0440]	−0.0364	12/28	0.0166	0.0029	0.0089	−0.0837
1.00	2	Nec (logit)	40	−0.358 [−0.882, 0.195]	−0.2383	16/24	0.2682	0.2075	0.2862	−0.4067
1.00	2	IoU	40	0.0930 [0.0046, 0.1818]	0.0262	24/13	0.0989	0.0477	0.0978	0.0946
0.50	0	Nec	40	−0.1138 [−0.2110, −0.0209]	−0.0074	15/25	0.1539	0.0277	0.0652	−0.0724
0.50	0	Nec (logit)	40	−0.713 [−1.475, 0.053]	−0.2565	15/25	0.1539	0.0803	0.1396	−0.7963
0.50	0	IoU	40	0.0325 [−0.0295, 0.0979]	0.0000	9/14	0.4049	0.3591	0.4505	0.0290
0.25	0	Nec	40	−0.0820 [−0.2127, 0.0382]	0.0062	21/19	0.8746	0.2168	0.2890	−0.0346
0.25	0	Nec (logit)	40	−0.981 [−2.141, 0.048]	0.0197	23/17	0.4296	0.0943	0.1509	−0.4595
0.25	0	IoU	40	−0.0256 [−0.0990, 0.0476]	0.0000	13/17	0.5847	0.5026	0.5291	−0.0257

citation loss, and it ends up citing *better*. The two arms are therefore not matched on localisation, and a contrast on the self-cited region $R(x)$ confounds mechanism with citation quality.

Two consequences, one methodological and one substantive.

Methodologically, the contrast we lead with is at the **ground-truth mask** $M^\star$ — the same region, the same operator and the same images for both arms, a region *neither arm chose*. That contrast cannot be moved by one arm citing better, and it is the same quantity Section H already uses for the FaithAlign arms for exactly this reason:

paired, global-cue − local-evidence, at $M^\star$	value	p
necessity (probability)	$-0.2764\ [-0.4033, -0.1507]$	$= 0.0003$
necessity (log-odds)	$-1.02\ [-2.11, 0.15]$	$= 0.0923$

with paired median -0.1954, sign test 29 of 40 negative ($p = 0.0064$), leave-two-out mean -0.2441 and Benjamini–Hochberg $q = 0.0012$ over the hybrid family. The absolute levels are what the hybrid corpus bought: the local-evidence arm's necessity at $M^\star$ is 0.3513 $[0.2402, 0.4665]$ against the global-cue arm's 0.0749 $[0.0247, 0.1993]$, a difference in probability units rather than the 10^{-4} the MagicBrush cell could reach.

The two scales agree on the distribution and can disagree on the mean. Unlike the MagicBrush cell of Section H.3.1, the local-evidence arm here is *not* saturated, so a difference of probabilities has real range and the raw scale is informative. The log-odds re-expression of the identical rows gives median -1.381 and sign 29 of 40 negative ($p = 0.0064$) — the same rows moving in the same direction — with a wider interval, because the log-odds transform amplifies the few rows where a probability sits near 0 or 1. We report both and lead with neither in isolation: the distributional evidence (median and sign test) is what agrees across scales, and it is what we would defend.

Substantively, the direction of the IoU difference is worse for IoU-only evaluation than a null would have been. The arm whose verdict is *not* caused by the manipulated region points at that region

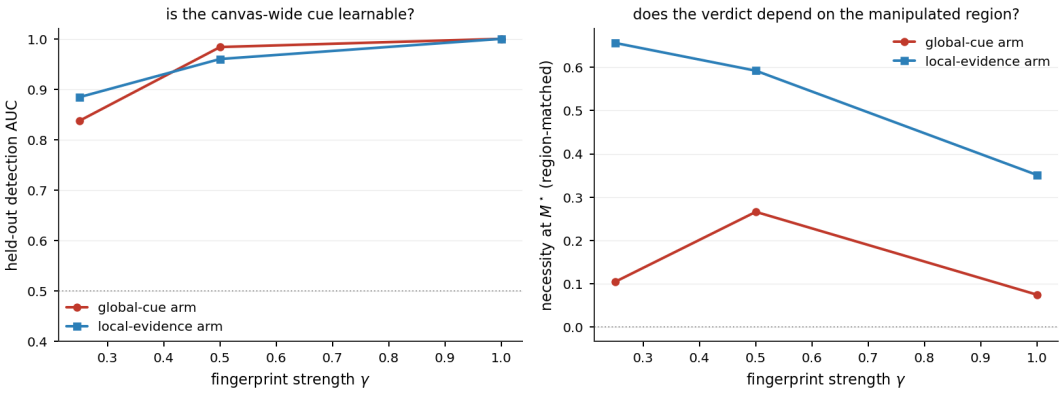


Figure 17: Edit-ID holdout; source independence unestablished. The fingerprint sweep.

better: an evaluator using localisation alone would rank the correct-but-non-causal arm *above* the causal one. That is a stronger form of the paper's claim than "IoU cannot separate them", and we did not design it — it fell out of the training dynamics.

At the self-cited region. For completeness, the contrast at each arm's own citation $R(x)$ is -0.0048 $[-0.0998, 0.0881]$ ($p = 0.9265$, $q = 0.9265$), paired median -0.0003, sign 20 of 40 positive ($p = 1.0000$), leave-two-out -0.0086; on the log-odds scale 0.40 $[-0.40, 1.41]$ ($p = 0.4841$). Because the citations differ, we do *not* read this as a mechanism contrast, and it is reported so that the region-matched number above cannot be mistaken for a selected one.

J.4 The jointly supervised detector also uses the global cue

The third arm, `detect_cite`, is trained the way the field trains: its positive class is the manipulated image and its negative class is the authentic one, so *both* cues are available and nothing tells it which to use. On this corpus the local evidence is exact and perfectly sufficient; a detector that used it would reach the same accuracy as one that used the fingerprint.

Its pattern is consistent with use of the fingerprint. Paired on the same 40 held-out rows, the necessity contrast against the global-cue arm is **unresolved** (0.0001 $[-0.0019, 0.0018]$, $p = 0.9414$); this is not an equivalence result. It is strongly different from the local-evidence arm's (-0.2763 $[-0.4036, -0.1496]$, $p = 0.0003$). In absolute terms it sits at 0.0750 $[0.0250, 0.1997]$ against the local arm's 0.3513 $[0.2402, 0.4665]$, at detection AUC 1.000 and mean $p(x) = 0.9997$. This sample does not resolve a difference between its verdict's dependence on the manipulated region and that of an arm which was *never shown a manipulated image at all*.

This is the paper's central pathology produced on demand rather than found: a detector trained normally, on a corpus where the local evidence is exact and sufficient, whose measured dependence on the manipulated region is weak — and which an evaluator using accuracy would call perfect and an evaluator using IoU would call well-localised (0.2600 $[0.1799, 0.3569]$; its contrast with the global-cue arm is unresolved, 0.0099 $[-0.0416, 0.0569]$, not equivalent). We did not design this arm to fail; we included it so that "the gain is just better detection" could be tested, and it is the arm that most directly indicts the way these systems are trained and evaluated. It also bounds the claim: what makes a detector unfaithful here is not a defect in the architecture or the objective but the *availability of a canvas-wide cue in the training distribution*, which is a property of the corpus that every generated-image detector is trained on (Wang et al., 2020; Frank et al., 2020).

Detection ranking is matched; response calibration is not. Both arms reach essentially perfect held-out detection on the audited pair (global-cue AUC 1.000, local-evidence 1.000), so the necessity difference is not a difference in accuracy. The untrained checkpoint is audited too and is itself instructive: it separates the two classes at AUC 0.999 with no training whatsoever, while its necessity at $M^\star$ is 0.0096 $[0.0038, 0.0323]$ — accurate and non-causal before any of our objectives are applied (Table 32).

Which scale to read, and what the log-odds scale costs. The two arms are trained on different class pairs and therefore need not sit at the same operating point, and Table 32 reports each arm's mean $p(x)$ so the reader can check. Where they differ, a difference of *probabilities* is partly a statement about the sigmoid and not about the pathway — exactly the defect Section H.3.1 documents on MagicBrush. Where the verdict is saturated the log-odds contrast is therefore the read with range, and the raw one is reported beside it; where the two disagree we say so rather than choosing.

Scale sensitivity across replications. Across the 3 seeds of §J.5, the log-odds re-expression of the self-cited necessity contrast is the least stable quantity in this paper: per-seed means 0.404, -0.819, -0.358 (across-seed sd 0.618), per-seed p 0.4841, $< 10^{-4}$, 0.2075, sign agreement 67%, every per-seed p below 0.05: **no** — 1 of 3 seeds *inverts the sign*. The region-matched contrast this section leads with is stable on the raw scale over the identical runs (-0.2764, -0.2772, -0.3206, sign agreement 100%, all per-seed p below 0.05: yes). The log-odds scale is preferred in this paper only where the raw scale has no range; it is not the more robust scale, and on this leg it is the less robust one.

Across the sweep. The measurement that is *not* entailed by the construction is whether the canvas-wide cue is learnable at all at a given strength. The global-cue arm's held-out AUC is 0.838, 0.984, 1.000 at $\gamma = 0.25, 0.50, 1.00$, and its necessity at $M^\star$ is 0.0266, 0.0527, 0.0293 against the local-evidence arm's 0.1086, 0.1665, 0.0341. Where the cue is too weak to be learned the arm is not a correct-but-non-causal detector, it is simply a bad one, and we do not read its necessity as evidence about faithfulness — which settings those are is itself the finding.

J.5 Replication across seeds

Across the replicated seed/split pairs, the fixed-region contrast is stable, while the self-cited contrast is not. Because the two arms' citations are not matched, the latter is an output-region comparison rather than a controlled mechanism contrast.

We re-ran the $\gamma = 1.00$ pair at 3 seeds, each with a freshly permuted train/eval split (distinct eval-split fingerprints: yes). **The region-matched contrast the section leads with replicates.** Necessity at $M^\star$, global-cue minus local-evidence, per seed: -0.2764, -0.2772, -0.3206 (mean -0.2914, across-seed sd 0.0253); per-seed permutation p: 0.0003, $< 10^{-4}$, $< 10^{-4}$; sign agreement 100%; every per-seed p below 0.05: yes. The *secondary* contrast at each arm's own citation, which the section already declines to read as a mechanism contrast because the citations are not matched, is correspondingly seed-sensitive: per seed -0.0048, -0.0875, -0.1134 (mean -0.0685, sd 0.0567), per-seed p 0.9265, 0.0006, 0.0029. That the primary leg is stable where the secondary one is not is the reason the primary leg is primary. On the log-odds scale: -1.023, -1.407, -1.082 (mean -1.171, sd 0.206). **The log-odds re-expression of that secondary leg is the least stable quantity in this replication, and we print it where we promote the scale rather than only in the results file**: per seed 0.404, -0.819, -0.358 (mean -0.258, sd 0.618), per-seed p 0.4841, $< 10^{-4}$, 0.2075, sign agreement 67%, every per-seed p below 0.05: **no** — 1 of 3 seeds *inverts the sign*. The log-odds scale is preferred in this paper only where the verdict is saturated and the raw scale therefore has no range; it is not more robust than the raw scale, and on this leg it is less so. global-cue arm: AUC per seed 1.000, 1.000, 1.000; necessity per seed 0.0293, -0.0000, 0.0071. local-evidence arm: AUC per seed 1.000, 1.000, 1.000; necessity per seed 0.0341, 0.0875, 0.1204.

The unit of replication is the (seed, split) pair, and the seed indexes both the initialisation and the permutation, so any spread reported here is a *conservative* one: it absorbs split-to-split variation as well as optimisation noise.

J.6 The same design with a real cue: the JPEG-75 replication

The fingerprint F_γ is ours, uniform by design, which makes it a clean instrument and unlike a real re-encode. The three-arm design is therefore re-run on a corpus whose canvas-wide cue is a JPEG-75 re-encode of the whole image — a real, non-uniform, frequency-domain artifact (Wang et al., 2020; Frank et al., 2020) — with the same 120 steps and citation supervision across 3 independently permuted seed/split pairs (120 eval rows total; fingerprints `f57c7fcf136a, 2790c1c24d73, e3d1f3a4cb98`). The seed-0 anchor has 40 eval rows, fingerprint `f57c7fcf136a`, and 69 minutes of training and audit. Detection is not response-matched: AUC is 0.997 for the global-cue arm,

0.866 for the local-evidence arm, and 1.000 for the arm trained the way the field trains. The comparison therefore demonstrates a learned dissociation, not a perfectly controlled mechanism swap.

Reading Table 34. Detection is reported both as ranking (AUC) and at the fixed 0.5 operating point; Brier and 10-bin ECE expose response-scale differences. IoU, fixed-region necessity and self-cited necessity appear for every arm rather than being transferred from the synthetic sweep.

Which arm is stable across seeds, and which is not. 3 seed/split pairs are 3 training replications, not 120 independent ones, so the per-seed table above is the primary report and any pooled statement is clustered by seed. Read that way the two *deliberately* trained arms hold their level: the local-evidence arm's fixed-region ceiling is 0.6058, 0.6534, 0.6615 across the three, a spread of 0.0557. The arm trained the way the field trains does not: 0.1712, 0.0048, 0.1926, a spread of 0.1878, which as a fraction of the local arm's dependence on the same seed is 0.28, 0.01, 0.29. So the replicated finding is the *contrast* between the two trained arms — 3/3 at the fixed region and 3/3 at the self-cited one — and not a proportion of local evidence that ordinary training selects. Any sentence in this paper that quotes the mixed arm's level quotes the seed it came from.

Historical exposed-cohort pooling, with the replication as the unit. This is a secondary descriptive summary of the original edit-ID holdouts, not source-clean confirmation. The strict sensitivity in Appendix P leaves seed 2 self-citation unresolved. Pooling the 120 audited rows as if they were independent would report 3 training runs as 120 experiments: rows read through one trained adapter are correlated, and the permuted splits share images (99 distinct images, 21 of them in two splits). We therefore pool the per-replication paired means with equal weight and use Student t on 2 degrees of freedom ($t = 4.30$), taking as the variance the larger of the between-replication variance of those means and their within-replication sampling variance, the latter including the covariance contributed by the shared images (`seed_clustered_pooled_contrast`; its null cases are pinned in `tests/test_seed_clustered_pooling.py`). Global-cue minus local-evidence, the pooled necessity gap is -0.6295 $[-0.8175, -0.4414]$ at $M^\star$ and -0.2933 $[-0.4696, -0.1170]$ at the self-cited box, negative in 3 of 3 and 3 of 3 replications; the pooled IoU gap is -0.0011 $[-0.1673, +0.1651]$, with the pooled sign in only 1 of 3. An exact sign-flip test whose unit is the replication cannot reach $p < 0.05$ with 3 replications (its smallest attainable two-sided p is 0.25), the t_2 interval does not overcome that precision limit or establish population coverage with only three fitted runs.

Region-matched. At the ground-truth mask $M^\star$ — the same region, operator and images for every arm — the global-cue arm's necessity is 0.0270 $[-0.0027, 0.1371]$ (median -0.0010), the local-evidence arm's 0.6058 $[0.4650, 0.7325]$ (median 0.8750) and the naive arm's 0.1712 $[0.0912, 0.2869]$ (median 0.0018). The paired global-minus-local contrast is -0.5788 $[-0.7137, -0.4377]$ ($p <$ 0.0001, BH $q = 0.0003$ over the family declared by the generator, median -0.8591, sign 36 of 40 negative, leave-two-out -0.5560, $n = 40$); the naive arm sits between them (+0.1442 $[+0.0430, +0.2549]$ against the global-cue arm, $p =$ 0.0092; -0.4346 $[-0.5883, -0.2740]$ against the local-evidence arm, $p <$ 0.0001).

Localisation. The arms' self-cited-box IoU contrast is unresolved: IoU 0.2708 $[0.1881, 0.3731]$ (median 0.1447) against 0.2667 $[0.1779, 0.3798]$ (median 0.0954), paired +0.0041 $[-0.0710, +0.0695]$ ($p =$ 0.9173, median +0.0000, sign 15 of 28 positive, $n = 40$). Its interval includes both zero and nontrivial changes, so we do not claim equivalence; the naive arm's point estimate is highest at 0.3364 $[0.2503, 0.4312]$. The IoU contrast is computed on every one of the 40 paired rows; an exact-zero IoU, which a box that misses the mask produces, is a measurement and is kept as one.

Self-cited. The trained arms box the edit inside their own narration, so this corpus also carries the narration-embedded self-citation test that the released system cannot (its explanation head emits no box): necessity at each arm's *own* cited region is 0.2974 $[0.1841, 0.4399]$ (median 0.0107) for the local-evidence arm against 0.0233 $[-0.0023, 0.1290]$ (median -0.0000) for the global-cue arm and 0.0501 $[0.0133, 0.1368]$ (median 0.0001) for the naive arm; paired global-minus-local -0.2740 $[-0.4004, -0.1550]$ ($p <$ 0.0001, median -0.0123, sign 30 of 40 negative, $n = 40$). The arm trained only on the canvas-wide cue cites a region with a much smaller restoration effect, while the observed IoU difference is unresolved. Two potential confounds remain: paste-back removes re-encoding as well as splice pixels, and the global-cue arm is audited outside its training distribution.

Table 34: **JPEG-75: ORIGINAL EDIT-ID HOLDOUTS, SOURCE-EXPOSED.**

Original edit-ID holdouts; source-exposed. Strict seed-2 self-citation remains unresolved.										
seed	arm	AUC	bal. acc.	$\bar{p}_1$	$\bar{p}_0$	Brier	ECE	IoU	$\mathrm{Nec}(M^\star)$	$\mathrm{Nec}(R)$
0	local	0.866	0.825	0.671	0.046	0.154	0.159	0.267	0.6058	0.2974
0	global	0.997	0.988	0.989	0.029	0.012	0.012	0.271	0.0270	0.0233
0	both	1.000	0.975	0.954	0.000	0.015	0.023	0.336	0.1712	0.0501
1	local	0.903	0.875	0.797	0.151	0.125	0.114	0.326	0.6534	0.3027
1	global	0.993	0.950	0.975	0.077	0.047	0.046	0.389	0.0000	0.0000
1	both	0.999	0.975	1.000	0.061	0.027	0.030	0.451	0.0048	0.0092
2	local	0.882	0.850	0.699	0.049	0.134	0.141	0.367	0.6615	0.3332
2	global	1.000	0.988	0.999	0.035	0.010	0.018	0.297	0.0054	0.0301
2	both	1.000	0.988	0.974	0.000	0.005	0.013	0.358	0.1926	0.0243

Their bias direction is not identified for arbitrary learned scores and interactions. It remains one backbone. Across the 3 source-exposed edit-ID seed/split pairs, global-minus-local fixed-region necessity is -0.5788, -0.6534, -0.6562, self-cited necessity is -0.2740, -0.3027, -0.3030, and IoU is +0.0041, +0.0630, -0.0705; the table, rather than a nonsignificant seed-0 test, is the original edit-ID replication record. The strict host/donor-unseen sensitivity in Appendix P uses the same fitted models and leaves seed 2 self-citation unresolved.

J.7 What the hybrid corpus does not settle

Three things, stated because they bound the claim.

The fingerprint is ours, not a generator's. F_γ is a fixed analytic field, not the artefact of a real generative model. It is uniform by design, which is exactly what makes it a clean instrument and exactly what makes it unlike a real re-encode, whose energy varies with content. The claim it supports is "when a canvas-wide sufficient cue exists, an accurate detector can show weak restoration dependence at the cited region, which necessity can reveal even when an IoU comparison is unresolved" — not a claim about how strong real generator fingerprints are. The MagicBrush arm is where a real fingerprint is measured, and it is the arm with no resolution; that trade is the honest state of the evidence.

Paste-back removes fingerprint as well as splice. A_R pastes authentic pixels, so inside R it removes the local manipulation *and* the fingerprint. A learned score need not respond monotonically to the amount of fingerprint removed. Without a positive-response sign assumption, this operation supplies neither a positive necessity floor nor an upper bound on region independence. It can enlarge, shrink or reverse a contrast through interactions.

The global-cue arm is audited off its training distribution. It never sees a spliced image during training: its positive class is the fingerprinted authentic host, whereas its audit input is the spliced target. Responses to that shift may have either sign. Holding the audit pairs identical controls the observed inputs, not the models' training distributions or internal responses. The measured contrasts remain operational observations on this construction.

The corpus is still a construction. Splice and copy-move with a feather kept strictly inside $M^\star$ are classical forensics, not the manipulations a deployed detector faces, and the fingerprint is added rather than inherited. What the corpus buys is not realism but the ability to know the answer in advance, which is the only thing that makes the positive-control protocol work at all.

K Limitations and scope

We state the weaknesses of this work plainly, because several of them bound what the results license.

The evaluation boundary: which claims rest on real models, and which do not. This is the most important limitation in the paper, so we state it first and without hedging. Section L.2 runs the *entire CF-Score audit* unchanged against three frozen open-weight VLMs (`llava:7b`,

`llava:13b`, `qwen2.5vl:7b`) served locally on 50 real DALL-E-2 MagicBrush edits each — 150 audited explanations, a complete balanced design. Necessity, specificity under a matched control, localisation consistency, the text-metric comparison, the control ablation, the inpainter-invariance test and the false-flip calibration are **real-model, real-data results**. They are also *black-box* results: nothing else was available.

Two experiments that require internal model access use a **locally hosted open-weight VLM**: Section H runs visual-token activation patching and the FaithAlign ablation on **Qwen2-VL-2B-Instruct** (2.21B parameters, float32, Apple MPS), audited by the identical CF-Score ledger. Those are real-VLM results *about that model*. Qwen2-VL-2B is a 2B model and an always-fake responder before training; it is not a deployed forensic system, and nothing measured on it transfers to `qwen2.5vl:7b`, to `llava`, or to FakeShield/FFAA. We never make that transfer and the section names the model in every claim.

What remains **synthetic-sandbox only**: the fifteen-detector study of Section L.1, which needs a large population whose causal structure we built, and the white-box half of the threat model (Section L.1). The sandbox study tests recovery of designed faithfulness. The primary validation is the positive control of Section D, which uses real images, real edits, real masks and the real audit path; the sandbox version is retained for its breadth of detector archetypes, not as the primary evidence. The adaptive attack is technically feasible on Qwen2-VL-2B — we verified that gradients w.r.t. `pixel_values` are populated — and was not run, so we claim nothing from it. A query-based *black-box* evasion attack likewise falls inside the served-model boundary and was simply not run, for budget: one SPSA trajectory costs 10^2–10^3 queries for a single image against 30 for a whole audit row. Consequently **no claim in Section L.1 is a claim about `llava`, `qwen2.5vl`, FakeShield, FFAA or any other deployed system.** The served-model constraint is still enforced in code: the served adapter returns `None` from every internals hook, so `supports_internals()` is false and the patching path cannot be entered against it; a unit test asserts it.

Coverage of released forensic systems. A gap between the population indicted and the population audited would be the single largest limitation of a paper like this one; what remains is a gap in *breadth*, not in kind, and it still deserves its own paragraph rather than a clause. The introduction motivates the problem with released explainable forensic VLMs — FakeShield, FFAA, ForgeryGPT, SIDA and their 2025–26 successors. The systems audited are (i) FakeShield v1-22B (DTE-FDM), a released explainable forensic VLM (Section E), and FFAA, a second (Section E′), (ii) three general-purpose VLMs *prompted* to do forensics, and (iii) a 2B model we train ourselves with the objective the literature uses (verdict loss plus imitation of a rationale that names and boxes the true edit) — the `detect_cite` arm of Section H.1.

Two released checkpoints are not the population. ForgeryGPT, SIDA and the 2025–26 systems named in the introduction are not audited; nothing measured on FakeShield or FFAA transfers to them by assumption.

Public checkpoints extend beyond the systems audited here. We probed for the weights and recorded the result (`scripts/probe_forensic_checkpoints.py`, `results/forensic_checkpoint_availability.json`). Under the identifiers the systems' own repositories link, `thu-huangzc/ffaa-mistral-7b`, `lingcco/fakeVLM`, `saberzl/SIDA-7B`, `saberzl/SIDA-13B`, `saberzl/SIDA-7B-description`, `saberzl/SIDA-13B-description` resolve HTTP 200, neither gated nor private, to an unauthenticated Hugging Face API request on 2026-09-05 (6 of 6 identifiers across 3 systems: FFAA, FakeVLM, SIDA); only ForgeryGPT returns 401 on every identifier we found (2 probed), a code the API returns both for a gated repository and for one that does not exist, so we do not claim its weights are gated — only that they do not resolve. What rules out a network or client fault is the control in the same run through the same client: 8 identifiers, including the FakeShield checkpoint this paper does audit, resolve 200, 401. The probe uses the namespaces linked by each repository's README. Thus **three members of the target population have public weights and two released checkpoints are audited under specified configurations** (FakeShield here, FFAA in Section E′), and what separates SIDA and FakeVLM from this paper is compute, not access. Three further caveats attach to the primary system we audited. Its explanation head emits no box (0% of 50 rows), so its citation comes from our post-hoc grounder while the served models cite their own boxes — a real asymmetry on every between-model citation comparison, which is why

the causal claims are also reported on the grounder-free regions where all four models are scored on identical pixels. MFLM, FakeShield's separate SAM-based localisation module, is loaded, gated and audited (Section E.6), which dissolves that asymmetry for the region the system *outputs* — the normalized intervention effect rises from 0.0920 $[0.0353, 0.1582]$ through our grounder to 0.2179 $[0.1158, 0.3289]$ through its own mask, so the grounder number is partly a measurement of our instrument. The ratio is not an additive allocation of internal evidence. It does not remove the ambiguity for the *narration*: the shuffled-text control does not establish that MFLM's mask is conditioned on what DTE-FDM wrote (+0.0758 $[-0.0285, +0.1829]$ against another image's narrative, $p = 0.5682$), so the re-audit is evidence about the region the system produces and not about the region its prose describes. And FakeShield's native answer is a paragraph whose first token is not a verdict, so the graded probability comes from a forced-choice suffix appended to FakeShield's own question; a second token family is reported alongside in the results file. A reader should read Section E as evidence about one deployed forensic system's *explanation head*, the served-model section as evidence about *prompted* general-purpose VLMs, and the white-box section as evidence about *one* 2B model trained the way the field trains one.

FFAA requires a face corpus with a retained authentic source. FFAA's weights are public, ungated and Apache-2.0 under the publishing author's namespace (`thu-huangzc/ffaa-mistral-7b`, HTTP 200; two other identifiers do not resolve), and we hold them locally. What the audit needs is the *corpus*. FFAA is a face-forgery model, so auditing it through this instrument needs a face corpus that *holds* the pre-edit source. The obvious candidate is CelebAMask-HQ paired with REFace swaps, whose identifiers encode the CelebA-HQ index of the image that was edited, so the authentic pre-edit image is recoverable in principle.

We certified that pairing before running anything on it, and it does not certify (`scripts/certify_reface_corpus.py`, `results/reface_corpus_certificate.json`). Over 60 pairs, paste-back at the face support (mean area 0.33 of canvas) is bit-exact outside that support on 0 of them; the outside-support difference runs to a maximum of 242 intensity levels with mean 7.20 and median 5.0, and 77.3% of outside-support pixels differ at all (84.6% of the whole canvas). REFace re-encodes the entire canvas, so the stored "authentic" image is not the fake with the swap undone: this corpus is MagicBrush-like rather than splice-like. A detector can earn AUC there from re-encode statistics that no local intervention reaches, and auditing FFAA on it would reproduce exactly the confound measured in Section E.3. REFace is therefore excluded from the local-evidence audit.

A locally composited face corpus meets the locality requirement. It stores the pre-edit source and composites the edit *locally*. `scripts/build_face_splice_corpus.py` builds one (`data_cache/celebsplice`) by the same construction and the same certificate as our splice corpus: CelebA-HQ images are aligned crops and CelebAMask-HQ ships per-part segmentation, so the central face region — eyes, brows, nose and mouth — of a donor can be composited into a host in `uint8` at matching coordinates. Re-verified from the written PNGs, it certifies: over 150 rows the maximum outside-mask difference is 0 and 0 rows change any pixel outside $M^\star$, so paste-back over the true mask returns the authentic image bit-for-bit. It is a *region-replacement* forgery, not a GAN swap — no identity blending, no colour harmonisation, no re-render — so a high AUC on it would not be evidence about wild face-swap detection; the manifest says so, and the faithfulness question conditions on detection in any case.

The other two preconditions are also settled, from files on disk and without materialising a weight (`scripts/probe_ffaa_loadability.py`, `results/ffaa_loadability.json`). FFAA's weights are held (15 GB, Apache-2.0). And its `config.json` names `LlavaLlamaForCausalLM` with `model_type="llava_llama"` — the same class DTE-FDM names — so we ran *this paper's existing rename*, `rightpixels.models.fakeshield.remap_key`, over its key set unchanged: it does apply, mapping 686 of 686 checkpoint tensors with 0 missing, 0 unmapped and 0 shape mismatches. The bitwise gate FakeShield passes is therefore runnable here too.

Released questions do not imply identical native pipelines. FakeShield's question is quoted *verbatim* from its released repository. FFAA's ten released user-question paraphrases are held verbatim (`data_cache/ffaa_release`, commit `a2be084`), but the historical adapter wrapped them in the generic LLaVA system line. The released FFAA `v1` conversation instead uses a specialized face-

forgery-analysis system line. Thus the historical FFAA rows do not reproduce its complete native prompt or its three-candidate MIDS selection. They remain observations of a released checkpoint under that explicitly specified wrapper. Section E′ reports the teacher-forced answer-field readout, absence of a native box, fixed paraphrase, part-name citation test with shuffled-narration control, the sufficiency direction, and the other nine paraphrases. ForgeryGPT's weights do not resolve under any identifier we found, and SIDA's are public, so its audit is compute rather than access.

What the positive control does and does not license. The oracle detectors are *privileged*: they are handed the source image and the ground-truth mask. That is legitimate for a positive control — the object under test is the metric, not the detector — and it is what makes "this verdict is caused by exactly these pixels" a fact rather than an assumption. But it means the ceiling of 0.729 is the score of an oracle, not of any achievable deployed system, and a real detector could be perfectly faithful and still fall short of it wherever the manipulation's evidence is partly global. The half-evidence arm (0.410 $[0.402, 0.416]$) is the honest illustration of that. Second, the control excludes triples whose ground-truth bounding box exceeds 35% of the canvas, by a rule fixed in advance, because an oversized box is clipped by the grounder and would cap the oracle's IoU for reasons unrelated to the metric; the excluded count is reported in the results file. Third, the blind-anomaly arm shows the metric measures *evidential*, not *computational*, necessity, and a reader who wants the latter should not read CF-Score as providing it.

The served-model sample is small and covers only two families. $n = 50$ pairs per model, 150 rows. The audit costs up to 30 forward passes per (model, image) and the serving host was shared with other workloads; we cut n rather than drop a condition, so the design is complete and balanced, the intervals are wide, and every table carries them. A fourth requested model, `llama3.2-vision:11b`, contributed only 2 exploratory rows (excluded from every reported number): it reserves 34 GB at its native context window on a server whose loaded-model limit is shared with other workloads, so it was evicted and reloaded per call at $\sim$110 s/row. That is an infrastructure fact about one box on one day and says nothing about the model's quality as a forensic detector, but it costs the study one architecture family — the reported set is two families plus a within-family scale contrast. Three further caveats belong here rather than in a footnote. Specificity under a *congruent* control is undefined for 26 of 150 rows because the cited region was too large for any disjoint translate to exist; those rows are recovered by the area-exact fragmented control of Section L.2 (coverage $0.827 \to 1.000$), at the cost of giving up the control's *shape* while holding its area exactly fixed — a substitution whose paired effect we measure (0.0143 $[-0.0055, 0.0379]$ ($p = 0.2165$, $n = 124$)) rather than assume. The MagicBrush "authentic" image is not bit-identical to the edited one outside the mask (DALL-E-2 re-encodes the whole canvas; median outside-mask MAD $\sim 1/255$), which is exactly why `qwen2.5vl:7b` retains 73% of its verdict separation after the true edit region is restored — necessity is bounded by the *local* component of the evidence and we report that bound alongside it — and Section I removes that bound entirely on a second corpus, where the two images *are* bit-identical outside the mask and the ceiling is the whole detection margin. And the explanation vocabulary of all three models is nearly constant ("inconsistent lighting" on 128 of 150 rows, and on $50/50$ for `llava:13b`), which limits what any text-overlap correlation can be asked to show — it is evidence for our claim that these metrics cannot separate systems, and simultaneously a reason the correlations we report have little dynamic range to work with.

Negative results from metric validation. The controls establish eight limits on interpretation. (i) Pooled $\mathrm{Nec} > \mathrm{Spec}$ held at an $n = 12$ pilot ($+0.068$, $p = 0.0085$) and did *not* hold at $n = 50$ ($+0.015$ $[-0.004, 0.041]$, $p = 0.20$); pooled necessity has an interval containing zero. More data moved the estimate toward zero, and the $n = 50$ figure is the one we report. (ii) The suspicion-matched control does not beat a random one — three independent replications of the same null — although the matcher demonstrably narrows the suspicion gap. (iii) Necessity is not inpainter-invariant on the served models: the sign inverts under both *classical* operators and rank agreement with paste-back is $\rho = +0.12$ and $+0.19$. (iv) But the *strong* form of (iii) — "inpainting is an invalid counterfactual operator" — is contradicted in turn by the positive control, where the same inpainters recover a known necessity to -0.013 with a 0% false-flip rate. The confound is a property of the detector, not of the operator. (v) Evidential dependence ("the cited region carries the evidence") differs from computational dependence ("the verdict is computed from the cited pixels"); the blind-anomaly arm scores 0.027 $[0.023, 0.032]$ while satisfying the second and not the

first, so the two are not interchangeable and CF-Score measures only the first. (vi) A gap against text overlap does not establish a composite-score advantage: localisation IoU alone reaches +0.728 $[+0.322, +0.908]$ against CF-Score's +0.915 $[+0.747, +0.973]$ on the sandbox population, and on the served rows $\rho(\mathrm{CF}, \mathrm{IoU}) = +0.990$[+0.985, +0.994], so the composite is close to a rescaled IoU wherever $\Delta\mathrm{NS} \approx 0$. The defensible claim is the global-cue contrast of Section D — a single comparison IoU provably cannot make — and not a general variance argument. (vii) The composite does not out-rank its own causal factor: necessity alone reaches +0.924 $[+0.772, +0.976]$ and the sensitivity-adjusted form +0.994 $[+0.982, +0.998]$ against CF-Score's +0.915 $[+0.747, +0.973]$ — the first two intervals are all but coextensive with the third, so the composite buys no ranking here. The product's justification is the separation of faithfulness from correctness, not a better ranking. (viii) The global-cue contrast of (vi) is *not* a contrast the closest prior work cannot make: EDCT's answer leg computed with our operator separates the same two arms completely (1.000 against 0.000, Section D). What survives is that the composite is *graded* where CCS saturates, and that it sees the incomplete-evidence cell CCS cannot. The contribution is therefore the operator and the validation protocol, not a universally superior scalar score.

Temperature sensitivity of the composite. $\mathrm{CF} = \sigma(\Delta\mathrm{NS}/T) \cdot \mathrm{IoU}$ has exactly one constant, fixed at $T{=}0.25$. The near-degeneracy with IoU on real populations follows directly from that choice, so the fair test is the sweep, recomputed from the saved $\Delta\mathrm{NS}$ and IoU at zero forward passes (Table 35, $T \in [0.02, 2.0]$). It does not rescue the composite. On the served MagicBrush rows $\rho(\mathrm{CF}, \mathrm{IoU})$ falls only from +0.987 at the shipped T to +0.883 at its best value ($T{=}0.02$) — no temperature makes the composite anything other than a rescaled IoU on a population where $\Delta\mathrm{NS} \approx 0$, because that is a fact about the population and not about T. On the positive control the shipped T is in fact *suboptimal* for the job the metric exists to do: at $T{=}0.02$ the arm-level Spearman against designed faithfulness is +0.840 over 8 arms, against +0.803 at $T{=}0.25$ — and necessity alone, which has no temperature, reaches +0.822. We keep $T{=}0.25$ because every shipped number uses it and re-tuning a constant to a ranking we also report would be exactly the kind of choice this paper's methodology section teaches a reader to distrust; but the sweep is the evidence for demoting the product, and it is reported in full.

Reading Table 35. No T removes the redundancy with IoU on a real population, and the shipped $T{=}0.25$ is not the best value for arm-level recovery of designed faithfulness.

Single-backbone training and control-selection sensitivity. Every trained arm is a LoRA on Qwen2-VL-2B-Instruct, at one rank and one step count. An incomplete SmolVLM-256M run (10 of 160 audit rows) is excluded. The corpus contrast in Section I, where the counterfactual term has no resolved effect, limits generalisation even within this backbone.

Activation patching is sensitive to control coverage. A congruent-only control retains 26, 11 and 15 of 55 rows, selecting for small citations. The coverage-complete audit uses an area-exact fragmented fallback and records control kind per row (165 rows versus 52 under the congruent-only rule). The FaithAlign cited-minus-control contrast is -0.0282 $[-0.0671, -0.0057]$ on the complete set, opposite in sign to 0.0306 $[0.0127, 0.0691]$ under congruent-only selection. Training increases dependence on visual tokens as a whole, but these results do not establish greater dependence on the cited region. Non-randomly excluding rows can reverse the estimated regional effect.

Limited coverage of learned inpainting. big-lama is evaluated in the full operator study (Section G, 100 triples, four oracle families). Two implemented operators are not evaluated: Stable-Diffusion-inpaint and MAT, whose adapters are implemented and gated in `rightpixels.counterfactual.adapters`. A denoising-diffusion inpainter writes a different trace from an FFC one, and since big-lama turned out to be *further* from a known necessity than either classical fill rather than closer, we do not extrapolate from it to SD: one learned operator is one point. We also note the name collision honestly: the operator whose registry key is `inpaint_diffusion` solves a heat equation and has nothing to do with denoising diffusion. The key is retained only because it is the key of every cached served-model response and of every shipped results file.

Grounding error limits narration-level conclusions. CF-Score is a function of $R(x)$, and a low score can mean "the explanation is unfaithful" or "we grounded it badly". We mitigate rather than

Table 35: **THE COMPOSITE'S ONE FREE CONSTANT.** CF recomputed from the saved ΔNS and IoU at each T (zero forward passes).

T	rho(CF, IoU) control	rho(CF, dNS) control	rho(CF, IoU) served	arm-level rho(CF, designed)
0.02	+0.902	+0.697	+0.883	+0.840
0.05	+0.898	+0.712	+0.923	+0.840
0.10	+0.894	+0.718	+0.957	+0.840
0.15	+0.893	+0.722	+0.973	+0.803
0.25	+0.893	+0.724	+0.987	+0.803
0.40	+0.906	+0.706	+0.995	+0.803
0.60	+0.925	+0.672	+0.998	+0.593
1.00	+0.946	+0.628	+0.999	+0.593
2.00	+0.967	+0.577	+1.000	+0.593

solve: three grounding settings evaluated side by side, a consensus grounder that reports pairwise agreement, and an explicit grounding-failure-rate column so unparseable or ungroundable citations are never silently dropped. The mitigation is not only a list of settings: Section E.4 runs the grounder against a *known* answer through the audit's own code path. It passes both gates that make the citation-null falsifiable at all — it beats a same-area random window when handed the true region in words (+0.079 $[+0.026, +0.146]$) and it moves when the named region moves (+0.143 $[+0.100, +0.206]$) — and it loses most of what it is handed: oracle text yields IoU 0.176 $[0.129, 0.238]$, 0.425 of the geometric ceiling, with the centroid inside the true mask on 40% of images, and we cannot attribute *any* of that recovery to the artifact class itself: naming it alone (0.091 $[0.058, 0.144]$) has an unresolved contrast with a random window (0.097 $[0.067, 0.132]$), which bounds its contribution without establishing that it is zero. Against the served models' own boxes it is at chance (0.098–0.132 against a floor of 0.119 $[0.099, 0.145]$; 0 of 3 contrasts significant). Abstract artifact language ("unnatural cheek lighting") is genuinely hard to map to a region, and this is what that costs.

Domain shift between authentic and manipulated images. On MagicBrush the question put to the detector is "has this image been manipulated?", and the two classes it must separate do not differ only inside the annotated mask — they arrive through different imaging chains. The authentic side is an MS-COCO photograph whose native size is 500×500, 1024×1024; the manipulated side is a DALL-E-2 inpainting output at 1024×1024, i.e. a *whole-canvas* generative re-render, and both are resampled onto a common 512×512 grid before anything is measured. The native sizes differ on 49 of 50 of the audited pairs. The consequence is measurable outside the mask alone: median outside-mask MAD $1.53/255$, minimum 0.05, and 0 pairs where it is zero. A detector can therefore obtain genuine, non-trivial AUC on this corpus from canvas-wide resampling and render statistics without attending to the edited region at all — and a *forensic* model prompted about manipulation is exactly the kind of system that would. This bears directly on our headline: the grounder-free ceiling effect is a real local sensitivity, but the *detection margin* it is measured against is not guaranteed to be local in origin, so "the verdict is anchored to the edit" should be read as "part of the verdict is" rather than "the verdict is, and only is". This is a confound in the corpus, not in the instrument, and Section E.3 is the reason it does not end the argument: the splice corpus is built by pasting real pixels into a real photograph, so both classes share one imaging chain exactly and the outside-mask difference is 0 by certificate. A result that survives there is not explained by render shift.

Automated adjudication is not independent human validation. The post-hoc grounder is assessed by (i) a positive control on the same images and code path, with oracle, decoy and same-area random arms; and (ii) an adjudicated subset of at least 20 explanation/region pairs with stimuli, labels, rationales and rater identity supplied for replication (Sections E.4 and E.7). The rater is the authoring system (a large language model), **not** an independent human. This is a model-based check, not independent human validation. The complete stimulus set and fixed decision rule permit human re-rating, which remains an external-validity limitation.

In that adjudication, on the 12 of 20 rows where the rater judged FakeShield's *text* to name the right place, our grounder recovered only IoU 0.048 $[0.012, 0.111]$, no better than the 0.093 $[0.025, 0.230]$ it recovered on the 8 rows judged wrong. A reader who concludes from this that the citation leg

measures our grounder as much as it measures the model is reading it correctly; what stops that reading from dissolving the section is that the grounder-free ceiling leg, which involves no citation at all, is where the positive result lives.

Multi-cause manipulations break the naive reading of necessity. An image that is fake for several reasons remains fake after one truthfully-cited region is restored, so a low Nec does not by itself imply an unfaithful citation. We measure the necessity of the *union* of cited regions and provide a single-artifact configuration for the clean test, and the sandbox exposes a multi-artifact fraction knob for exactly this ablation. Both numbers must be reported together; reporting only the single-region value would overstate unfaithfulness.

Paste-back's identity requires a source; partial restoration still has interaction ambiguity. Proposition 2 applies to constructed corpora, which are exactly the least ecologically valid ones. On in-the-wild data (IMD2020) there is no pre-edit source, restoration must be generative, and the argument rests on the region-conditional calibration and on the inpainter-invariance of the necessity ranking. That is a weaker guarantee, and the constructed↔wild gap should be reported explicitly rather than averaged over. Even with a held source, the proposition proves an identity on authentic inputs, not additivity or absence of boundary and context interactions for partial restorations of a manipulated input.

The suspicion-matched control is not vindicated. We give three answers to the circularity objection (Section C) and the sensitivity-adjusted necessity is empirically the strongest sub-score in the sandbox, but matching has never once beaten a random control on a downstream quantity, in three attempts, and we have no proof that it cannot deflate Spec under some detector we have not built. The honest statement is that the residual coupling is measured and reported (β_1, R^2 of Eq. 9), not that it is zero — and that on the evidence so far, the control's job is to subtract the *operator's* trace, which is why it matters for generative counterfactuals and is close to a formality for paste-back.

The security result is a correlation with a fitted bound, not a theorem. We deliberately do not prove that evasion cost decreases in $\Delta\mathrm{NS}$ under a decoupled-heads model: that theorem assumes what it concludes. What we have is a measured monotone relationship, significant after controlling for the detector's decision margin, and a quantile-regression certificate validated on held-out examples. At the scale reported here the sample is small ($n = 16$ paired points, 6 held out), and the raw correlation is not significant on its own. With only 6 held-out observations, empirical coverage cannot establish reliable calibration; the sandbox study includes an under-covering certificate. These are reported as they are.

Global-cue detectors admit no faithful region citation at all. For a model driven by a non-localisable fingerprint, no region counterfactual can be large, and CF-Score will be near zero regardless of which region is cited. We treat this as a positive finding – such a model's region narratives are unfaithful *by construction* and should not be shown to moderators as evidence – but it does mean CF-Score cannot distinguish among such models, and we report it rather than presenting it as a metric failure.

FaithAlign is demonstrated on a 2B VLM, not a 7B one, and not with the full objective of Eq. 14. Section H trains LoRA adapters on Qwen2-VL-2B-Instruct with real gradients, and Section I repeats the whole ablation on a second corpus. What remains open is scale and *objective coverage*. Of the five terms in Eq. 14, three are trained on the real VLM ($\mathcal{L}_{\mathrm{NLL}}$, $\mathcal{L}_{\mathrm{ground}}$ and $\mathcal{L}_{\mathrm{cf}}$); $\mathcal{L}_{\mathrm{attr}}$, which uses second-order autograd through the verdict logit for attribution alignment, and $\mathcal{L}_{\mathrm{size}}$ are **not** trained on any real VLM in this paper and their contribution is therefore untested at that scale. We also report two *forms* of $\mathcal{L}_{\mathrm{cf}}$: the softmax over restoration effects that the shipped `faithalign` arm trains, and the hinge of Eq. 15 that the Method section prints, trained here as a separate arm precisely because a reader following Eq. 15 would otherwise be reading a null against an objective we did not run.

The text-metric baselines are partly offline stand-ins, and the one served judge we ran was degenerate. BLEU-4, ROUGE-L and CIDEr are exact. BERTScore uses BERTScore's algorithm over deterministic hashed character-n-gram embeddings rather than a contextual encoder,

and the offline judge is a deterministic rubric; real backends are wired behind the download flag. On the real run we additionally scored a genuinely served text-only judge (`qwen2.5:7b`), and it was *degenerate* — NO on $150/150$ explanations with a dynamic range of 3.6×10^{-5} — so the one correlation Benjamini–Hochberg keeps is an artifact of ranking floating-point noise, which our `metric_degeneracy` diagnostic flags and carries into the BH record. This degeneracy prevents a meaningful graded comparison. It means we do *not* have a non-degenerate served LLM judge as a baseline, which is a real gap: our qualitative claim (that text-only measures cannot see causal faithfulness) follows from what those measures are functions of, but the specific correlation values would change with a working judge, and the served-judge row should be read as missing rather than as evidence either way.

The hybrid corpus buys resolution by manufacturing the confound. Section J's canvas-wide fingerprint F_γ is a fixed analytic field, not the artefact of a real generative model: it has uniform energy by design, which is exactly what makes it a clean instrument and exactly what makes it unlike a real re-encode, whose energy varies with content. The claim it supports is conditional — "*when* a canvas-wide sufficient cue exists, an accurate detector's verdict can be independent of the region it cites, and necessity detects that while IoU cannot" — and it is not a claim about how strong real generator fingerprints are. The JPEG-75 construction adds a real re-encoding cue, while MagicBrush illustrates the low-resolution case. Two further caveats: the paste-back operator removes the fingerprint as well as the splice inside the cited region, which can change a learned score in either direction; no sign assumption establishes a positive floor or a guaranteed attenuation of the contrast; and $\gamma = 0$ is excluded from the sweep because the global-cue arm is undefined there, so the sweep starts at the smallest γ at which the experiment exists rather than at the origin.

Multiple-comparison families and post hoc scale selection. The white-box family in Section H.3.1 contains both the raw and the log-odds version of several contrasts. Those are monotone re-expressions of the same stored probabilities, not independent tests; Benjamini–Hochberg is valid under the positive dependence they exhibit, but the family size overstates how many questions were asked. We report the original 18-test family, our 35-test family and their 53-test union precisely so the reader can see that the verdict on the raw-probability contrast changes with the family and the verdict on the log-odds contrast does not. Choosing to lead with the log-odds scale is a decision we made *after* seeing that it separates better, and although the argument for it (saturation, Section C.2) is mechanical and was stated before the correction was run, a reader is entitled to treat it as one degree of freedom we used.

No white-box adaptive-attack evaluation. Pixel gradients are available on this machine — `pixel_values.grad` is populated for Qwen2-VL-2B under `transformers` on Apple MPS, verified directly — so a white-box adaptive adversary against the audit is feasible and was not executed. The attack material that does exist (Appendix L.1) is sandbox-only, $n = 16$, with a certificate that under-covers. It provides no robustness guarantee for a released VLM.

The JPEG-recompressed corpus variant was built and audited. Section I discloses the hard-paste construction of the local-evidence corpus and ships two further variants (feathered strictly inside the mask, and JPEG-75). The feathered variant carries the hybrid results. The JPEG-75 variant has a measured outside-mask residual (median 0.102 levels, maximum 68.7) and a completed served-model audit on the same 50 host photographs (Table 30). Its normalized intervention ratio is reported only because its ratio-of-means and mean-of-row-ratios agree in sign; the loss of exact locality is reported beside it. Table 30 reports this completed corpus audit.

K.1 Sensitivity to corpus, estimator and instrument

The following results delimit the interpretation of the audit.

Corpus and statistical power. On MagicBrush, FakeShield's cited-region effect is unresolved (ΔNS 0.000 $[-0.024, 0.027]$); the design could not resolve an effect smaller than 80% of the detection margin. On the certified splice/copy-move corpus, the estimate is 0.0233 $[0.0050, 0.0563]$ at $n = 50$ and 0.0198 $[0.0104, 0.0331]$ at $n = 150$. On feathered seams it is +0.0041 $[-0.0162, +0.0235]$ at $n = 50$, with an interval including both zero and the hard-paste estimate.

Thus the grounder-based citation effect is established on the hard-paste construction, but unresolved on the feathered one.

Estimator sensitivity. At $n = 50$, the mean intervals and rank tests support a positive citation effect, but the paired permutation test does not ($p = 0.059$). At $n = 150$, the three tests and two mean intervals support a positive effect, while the median interval still touches zero (+0.0010 $[+0.0000, +0.0065]$). Mean effects should therefore not be read as effects shared by most images.

Region instrument. FakeShield's post-hoc text grounder is at chance against a served model's own box. The normalized intervention effect is 0.2179 $[0.1158, 0.3289]$ through the native MFLM mask versus 0.0920 $[0.0353, 0.1582]$ through that grounder. Neither ratio is an additive allocation of internal evidence, and MFLM's mask is not shown to depend on the narration.

Control coverage. On white-box arms, the congruent-only control non-randomly excludes large citations and yields a FaithAlign contrast of 0.0306 $[0.0127, 0.0691]$. With the area-exact fragmented fallback, it is -0.0282 $[-0.0671, -0.0057]$, with the opposite sign. The complete-cohort estimate is primary.

Composite-score redundancy. The composite CF closely tracks IoU on the evaluated real-data populations (Appendix O). Necessity and localisation are therefore reported separately.

Model coverage. Under repository-linked identifiers, SIDA, FFAA and FakeVLM have public, ungated checkpoints (`thu-huangzc/ffaa-mistral-7b`, `lingcco/fakeVLM`, `saberzl/SIDA-7B`, `saberzl/SIDA-13B`, `saberzl/SIDA-7B-description`, `saberzl/SIDA-13B-description`). Limited evaluation breadth is not evidence that the remaining checkpoints are inaccessible.

Exact-zero outcomes. The JPEG-75 paired IoU analysis retains all 40 pairs, including exact zeros, yielding +0.0041 $[-0.0710, +0.0695]$ ($p = 0.9173$). The interval includes zero and does not establish equivalence.

Replication units. The white-box replication in Section H uses permuted train/evaluation splits. The aggregator rejects identical split fingerprints as separate replications; repeated runs on one split are not independent seed/split evidence.

Experiments beyond the current operational claim. A stronger unseen-source learned claim needs a pre-training split of whole joint source/donor components and a prespecified gold/self-cited comparison. Broader narration validation needs another source-holding edit family and a more detailed region vocabulary that can falsify artifact descriptions beyond three-group naming. A model-faithfulness ranking needs matched corpus, region instrument and readout. Direct native-decision interventions, and FFAA's full answer-selection path, are separate readouts to evaluate. None was performed for this revision; none is inferred from the present elicited-score effects.

L SUPPLEMENTARY SANDBOX AND SERVED-MODEL ANALYSES

These supplementary analyses test the instrument on synthetic detectors with known causal structure and on general-purpose served VLMs. They do not support claims about released forensic systems, which are evaluated in Appendices E and E′. Complete protocols and per-row results are supplied in `notes/NOTES.md` and `results/`.

L.1 THE SYNTHETIC SANDBOX AND ITS FOUR EXPERIMENTS (E1–E4)

Authentic images were generated procedurally with a $1/f$-like spectrum, explicit object boundaries and per-image sensor-noise and colour-gain parameters, and manipulated by splice, copy-move and inpaint-style edits whose ground-truth support is the actual difference map rather than a nominal rectangle. Fifteen narrating detectors of known composition were audited through the same code path as every released model here. Its purpose was never external validity: a faithfulness metric

cannot be checked on models whose faithfulness is unknown, and this is the population where it is known by construction. The primary positive control in Appendix D uses real images; the sandbox provides a broader range of detector archetypes.

Four experiments ran on it. **E1** scored text-overlap quality against the causal contrast and found no association, the result the served-model audit and then the FakeShield audit reproduced on real data. **E2** patched visual token activations to separate the region a detector computes from the region it cites. **E3** trained the FaithAlign objective and found it buys localisation and no faithfulness, a negative result that survived to the released-model scale and is why the objective is not among this paper's claimed contributions. **E4** measured explanation-preserving evasion cost against the causal contrast and fitted a lower-quantile budget certificate; it is sandbox-only at $n = 16$ with a certificate that under-covers, it is not a claimed contribution, and Section K states plainly that the white-box adaptive attack was never run.

Two sandbox sub-studies are cited elsewhere for their *method* rather than their numbers. The *control ablation* compared suspicion-matched, random and anti-matched control regions; the matched sampler lands closer to the cited region's suspicion than uniform sampling, but under paste-back specificity is near zero for every detector, so the ablation does not separate detectors and we report the unresolved contrasts. The *inpainter-invariance* study swapped the counterfactual operator while holding the detector fixed; its conclusion — that the operator, not the metric built on it, determines the answer — is reproduced on real images with a learned inpainter in Appendix G, which provides the real-data evidence.

L.2 The served-VLM audit on MagicBrush

Three general-purpose chat VLMs served locally (`llava:7b`, `llava:13b`, `qwen2.5vl:7b`) were audited on 50 MagicBrush dev pairs each. Three limitations determine its scope:

- **The corpus could not carry the claim.** MagicBrush edits are whole-canvas DALL-E-2 re-renders, so paste-back at the cited region is not the identity outside the annotated mask and most of the detectable evidence lies where no local intervention reaches. Appendix I measures that directly and Appendix E therefore runs the identical instrument on a certified bit-exact corpus.
- **The control was a selection effect.** The congruent suspicion-matched control is geometrically impossible above roughly a quarter of the image area, so the audit was conditioning on small citations. Replacing it with an area-exact fragmented control raised coverage from 0.827 to 1.000 on the served rows and changed two estimated contrasts. The coverage-complete rule is primary; Appendix K reports its sensitivity.
- **Faithfulness is undefined where detection fails.** Stratifying on whether a model detected the manipulation at all left `llava:7b` with an empty detected stratum, which is reported rather than pooled away. The stratification rule is applied to every audit here.

Text-based baselines give an unresolved result: on these models no text-overlap or judge metric predicted the causal contrast, with every interval covering zero and nothing rejected under Benjamini–Hochberg. The released system reproduces it (Appendix E.11), and the condensed served-model contrast the main text uses is Appendix F.

M The real corpus: MagicBrush preparation and alignment gate

Section L.2 is measured on MagicBrush (Zhang et al., 2023) dev (`osunlp/MagicBrush`, declared CC-BY-4.0; COCO source photographs, edits produced with DALL-E 2 by the MagicBrush authors). Each record is a triple (authentic source, real edit, binary edit mask), rescaled so the long edge is 512 px on the grid the mask is defined on — bilinear for images, nearest for the mask after binarisation at native resolution.

Alignment gate. A source/target/mask triple is only usable as a counterfactual if the annotated mask really is where the two images differ. We compute the per-pixel mean absolute difference between source and target inside and outside the mask, and keep a record only if the inside MAD

exceeds $5/255$ and the inside/outside ratio exceeds 3. Of 264 dev records, 262 pass (2 dropped as *edit-too-faint*); the median inside/outside ratio is 60.

Two properties of this corpus that the audit must not paper over. First, MagicBrush is *multi-turn*: for turn ≥ 2 the "authentic" side of a pair is itself an earlier edit, and 127 of the 262 records are of that kind. We therefore restrict to turn-1 pairs, leaving 135; the mask-area window $[0.01, 0.6]$ leaves the 126 from which the audited 50 are drawn. Second, the "authentic" image is **not** bit-identical to the edited one outside the mask: DALL-E 2 re-encodes the whole canvas, and the median outside-mask MAD is $\sim 1/255$. Small, but evidently not invisible to a detector — this is precisely the effect that leaves `qwen2.5vl:7b` with 73% of its verdict separation intact after the true edit region has been restored (Section L.2). Paste-back inside a region is still exact by construction, so Proposition 2 is unaffected; what the re-encode bounds is the *ceiling* on necessity, which is why we report Nec against the restore-the-true-mask value rather than against 1.

N Implementation notes

Commands. The reproducibility statement describes each stage in words; these are the targets of the released `Makefile` that run them. Offline, no weights: `make test` (the suite), `make smoke` (the sandbox pipeline end to end), `make paper` (this document), `make validate` (the positive control, the classical-operator study and the stratification, given the MagicBrush download), `make positive-control-lama` (the *learned* operator leg, big-lama; CPU only, no GPU, $\sim$29 wall-clock minutes at $n = 100$ as measured, and it needs the big-lama torchscript download), `make citation-robustness` (every citation-leg estimator from stored rows), `make mflm-coverage` (the native-mask ratio against its coverage of $M^\star$, from stored rows), `make hybrid-analyze` and `make hybrid-realcue-analyze` (the hybrid aggregates and macros), `make final-tables` (only the current load-bearing tables from frozen rows), and `make claim-manifest` (claim/cohort/instrument hashes and commands), `make checkpoint-availability` (the weight probe), `make paper-figs` (the main-text figures), and `make supplement` (this archive, rebuilt; it re-reads every text member out of the zip and deletes the archive rather than shipping it if one carries an author-identifying path. It is built without the trained weight files, which no reported number is recomputed from; each adapter's configuration stays in, so the LoRA rank and adapted-module scope remain checkable). With the FakeShield checkpoint on disk: `make fakeshield` (the MagicBrush audit), `make fakeshield-splice` (the certified corpus at $n = 50$), `make fakeshield-splice150` and `make fakeshield-splice150-certify` (the extension and its bit-exactness certificate), `make fakeshield-splice-feather` (the feathered seam), `make mflm` (the system's own localiser: equivalence check, exploratory acceptance gate, audit and paired comparison), `make mflm-feather` (the matched construction replication), `make mflm-sufficiency` (necessity and sufficiency on the same masks), and `make fakeshield-readout` (native-prose/verbalizer agreement), and `make ffaa-parts` (the second system with shuffled narration and sufficiency). With a local Ollama server: `make real-audit`. With `torch`, `transformers`, `peft` and a GPU/MPS device: `make whitebox-faithalign`, `make whitebox-patching`, `make whitebox-splits`, `make whitebox-splice`, `make whitebox-globalcue`, `make hybrid`, and `make hybrid-realcue-seeds`. A re-audit of the certified corpus at $n = 50$ through the refactored FakeShield loader is reproduced by `scripts/queues/round10_loader_audit.sh`, which writes to its own `*_postrefactor` results files separately from the baseline files; it costs 1482 forward passes and about 57 minutes on one MPS device, and runs three stages — the audit, a cell-by-cell certificate against the baseline rows, and the primary contrasts recomputed from the new rows. Each stage's sentinel is minted on a post-condition over the written rows rather than on the process exit code, and none of those post-conditions requires the two runs to agree: a stage that demanded agreement could not report a disagreement.

Cost. The audit's dominant cost is the occlusion sweep, G^2 forward passes per (model, image). Two things make a 15-model audit cost barely more than a one-model audit: the occlusion lattice and the lattice-cell restorations are model-independent and are built once per image, and all detectors share a cue cache keyed by image content, so only the first model pays extraction. Candidate

suspicion during control search is read from the already-computed sensitivity map rather than re-occluded, turning an $O(n_{\text{cand}})$ sweep into $O(1)$.

Offline counterfactual operators. Three of the four comparison operators are classical; the fourth is learned and was run. The classical three are a red-black SOR solution of Laplace's equation inside the hole (a Navier-Stokes/Telea-family smooth fill, Telea, 2004), a PatchMatch-flavoured exemplar fill, and the same Laplace fill with sensor noise re-matched to the surrounding ring; the learned one is big-lama, below. (The registry key of the first is `inpaint_diffusion`, which is a name collision with denoising diffusion and means nothing of the sort; the key is kept only because it is the key of every cached response and every shipped results file, and `LaplaceInpainter` is the preferred alias in code.) They span different trace signatures on purpose. The LaMa adapter (Suvorov et al., 2022) is evaluated: `simple-lama-inpainting` supplies the released big-lama torchscript, and one obstacle is worth recording because it is silent — the checkpoint was traced on CUDA and the package's constructor calls `torch.jit.load` with no `map_location`, so on a CUDA-free machine it dies inside an ATen dispatch rather than with a useful message. We load it ourselves with `map_location`, pad to a multiple of 8 and crop back; it costs 1.4 s per 512×512 image on CPU, which is why the whole operator study runs without touching the GPU. Adapters for Stable-Diffusion inpainting (Rombach et al., 2022) and MAT remain implemented, gated and **not run**. One numerical constraint is essential: successive over-relaxation with $\omega > 1$ is stable only under Gauss–Seidel updates. Applying the same ω to a simultaneous Jacobi sweep diverges to a saturated checkerboard, which a forensic detector then flags with probability 1 – producing a "false-flip rate of 1.0" that looks like a finding about inpainting confounds and is actually a numerical bug.

FaithAlign architecture. A plain RGB CNN trained on a few hundred images does not discover noise-level inconsistency: the signal is orders of magnitude below the content. As in SRM and constrained-convolution forensic networks (Fridrich & Kodovsky, 2012; Bayar & Stamm, 2016), the trainable stack sits on a *fixed* high-pass front-end (residual, local residual standard deviation, Laplacian). The learnable part is free to ignore it and learn the global shortcut instead – which is the point, since a model that cannot express the failure mode is not a fair testbed for a metric that claims to detect it.

Offline metric backends. BLEU-4, ROUGE-L and CIDEr are implemented to their standard definitions. Two deviations are recorded here rather than buried. (i) CIDEr's IDF term is exactly zero when every reference n-gram occurs in every reference, which makes the metric return 0 for all candidates on a small or homogeneous corpus; we fall back to a smoothed $\log((N+1)/(d_f+1))+1$ in that regime and expose the strict form as a flag. (ii) BERTScore uses BERTScore's *algorithm* (greedy cosine alignment, precision/recall harmonic mean) over deterministic hashed character-n-gram embeddings when no encoder weights are available; a real contextual encoder is wired behind the download flag. The LLM judge is likewise a deterministic rubric offline and an API call when enabled. In all three cases the paper's qualitative claim follows from these being functions of *text alone*, which no backend changes.

Reproducibility. Figure colours are indexed by a sorted category list rather than by `hash()`, which Python salts per process; results files encode non-finite floats as `null` rather than `NaN` so they are valid JSON; and every results file embeds the resolved device, the full configuration and the wall time.

O LIMITATIONS OF A COMPOSITE FAITHFULNESS SCORE

We evaluate a candidate composite that multiplies the causal and localisation factors,

$$\mathrm{CF}(x) \;=\; \sigma(\Delta\mathrm{NS}(x)/T)\cdot L(x), \tag{21}$$

with T a temperature setting the dynamic range (ΔNS is a difference of probabilities and lives in $[-1,1]$). The following diagnostics motivate reporting its constituent quantities separately.

What we measured.

- **It is a rescaled IoU on every real population we tested.** $\rho(\mathrm{CF}, \mathrm{IoU}) = +0.990$ $[+0.985, +0.994]$ on served rows and +0.961 $[+0.955, +0.967]$ pooled over the positive-control arms, against $\rho(\mathrm{CF}, \Delta\mathrm{NS}) = +0.120$ $[-0.058, +0.291]$.
- **No temperature removes that.** A sweep of the metric's one free constant over $T \in [0.02, 2]$ never drops $\rho(\mathrm{CF}, \mathrm{IoU})$ below +0.883 (Table 35). This is a fact about the population — $\Delta\mathrm{NS} \approx 0$ almost everywhere — and not about the constant, which is why it cannot be tuned away.
- **Its own causal factor out-ranks it** at recovering designed faithfulness on the positive control: necessity alone reaches +0.924 $[+0.772, +0.976]$ and the sensitivity-adjusted form +0.994 $[+0.982, +0.998]$ against the composite's +0.915 $[+0.747, +0.973]$.
- **On the trained-network cell built to decide it, it is silent.** Two Qwen2-VL-2B LoRA arms with matched citations and matched detection are separated by necessity ($p = 0.011$) and not by the composite (-0.0133 $[-0.0465, +0.0189]$, $p = 0.459$): with $\Delta\mathrm{NS} \approx 0$ on both arms the sigmoid is pinned at $\frac{1}{2}$ and CF inherits IoU's null.

Why it fails, stated mechanically. The composite multiplies a factor that is near-constant on real populations by a factor that varies, so it inherits the varying one's ordering. That is not a bug in the constant T; it is what a product does when one factor has no variance. Any aggregate of a causal and a localisation term will do the same thing on a population where the causal term is compressed — which is precisely the population that motivated the metric. The lesson we would carry forward is that an aggregate should not be proposed until the marginal variance of each factor has been measured on the target population, and that a metric which correlates at $\rho > 0.98$ with one of its inputs should be reported as that input.

Reporting convention. The main text reports Nec and L_{gt} separately. Supplementary tables and results files labelled "CF" evaluate Eq. 21; these values permit reproduction of the composite-score diagnostics.

P SOURCE EXPOSURE, SCORE SCALE, AND PAIRED ROBUSTNESS

These are exploratory sensitivity analyses of stored inference scores. The specification in `analysis_specs/review_cycle1.md` was written after the original outcomes were inspected and the source-overlap problem was identified, but before these diagnostic outcomes were computed. It is not a preregistration or an untouched confirmation set. Every original result remains in its original cohort; exclusions below are functions of provenance, not scores. The diagnostic result manifest, `results/review_cycle1.json`, hashes the specification, source ledger, stage results and their input rows; `make review-diagnostics` reconstructs the tables without model calls.

P.1 EDIT-ID HOLDOUT AND SOURCE-IMAGE EXPOSURE

The JPEG corpus contains both a splice and a copy-move from some authentic photographs. Its original train/evaluation partition was over edit IDs, not source images. In seeds 0/1/2, 12/22/11 of 40 evaluation edits share a host photograph with training. We report the original cohorts, a host-unseen subset, and a stricter subset excluding any evaluation host *or donor* that appears as a training host *or donor*. The latter retains 18/10/15 pairs. All arms use exactly the same retained IDs within a cohort. Decoded RGB source hashes verify that repeated host IDs really refer to the same pixels and that different stored host IDs do not alias. Donor IDs come from the construction manifest; 23 donor photographs have no full source stored in this reduced cache, so their identity is verified by provenance ID, not independently by pixels. The released ledger records this distinction.

This exclusion cannot retroactively make the original training design source-disjoint. It estimates effects on the retained examples, which are smaller and have a different edit-family mix. Training leakage in the full cohort is a limitation even when the restricted effect has the same sign. The synthetic-cue and training-ablation tables likewise describe edit-ID holdout unless they explicitly establish source independence; we make no source-generalisation claim from them.

P.2 A MONOTONE-INVARIANT INTERVENTION DIAGNOSTIC

For one trained arm, let $t_i = p(x_i)$, $r_i = p(A_{R_i}(x_i))$ and $s_i = p(x_{\mathrm{src},i})$. With $h(a,b) = \mathbf{1}[a > b] + \frac{1}{2}\mathbf{1}[a = b]$, define

$$D_R = \frac{1}{n^2}\sum_{i,j}\{h(t_i, s_j) - h(r_i, s_j)\} = \mathrm{AUC}(t, s) - \mathrm{AUC}(r, s). \tag{22}$$

Both comparisons use the same reference negatives. For every strictly increasing q, $h(q(a), q(b)) = h(a, b)$, proving invariance of D_R to a common monotone recalibration of that arm's three score sets. No threshold, temperature or operating point is fitted using evaluation outcomes. This diagnostic is an AUC decrease, not a probability difference, so its units and interpretation differ from raw necessity. It can be negative; saturation in ranking can also hide score changes.

For each of 10,000 bootstrap draws (seed 20260913), we resample source-image clusters, retaining all edits of a selected source and every arm/state. If w_i is the multiplicity of row i, the draw is $\sum_{ij} w_i w_j\{h(t_i, s_j) - h(r_i, s_j)\}/(\sum_i w_i)^2$: the negative reference distribution is resampled together with the positives. The contrast is always global-cue minus local-evidence, with paired cluster weights. The percentile intervals below are pointwise descriptive intervals, not multiplicity-adjusted confirmatory tests. Reused donors can induce further dependence; source clustering does not claim to model all construction dependencies. The three training replications remain only three units, and their exact two-sided sign test cannot reach $p < 0.05$. We do not pool their evaluation rows as independent observations.

Table 36: **Ranking sensitivity to restoration.** Global-minus-local D_R and source-clustered 95% intervals; all arms share each cohort's provenance exclusions.

cohort	seed	pairs	sources	ΔAUC gap at M^*	ΔAUC gap at own region
original	0	40	36	-0.469 $[-0.581, -0.375]$	-0.170 $[-0.258, -0.094]$
original	1	40	39	-0.484 $[-0.560, -0.422]$	-0.212 $[-0.308, -0.124]$
original	2	40	36	-0.435 $[-0.517, -0.360]$	-0.201 $[-0.295, -0.104]$
host unseen	0	28	24	-0.519 $[-0.646, -0.413]$	-0.182 $[-0.304, -0.085]$
host unseen	1	18	17	-0.506 $[-0.640, -0.398]$	-0.194 $[-0.333, -0.080]$
host unseen	2	29	25	-0.459 $[-0.556, -0.378]$	-0.203 $[-0.310, -0.102]$
host/donor unseen	0	18	16	-0.562 $[-0.730, -0.418]$	-0.176 $[-0.329, -0.052]$
host/donor unseen	1	10	9	-0.500 $[-0.711, -0.321]$	-0.200 $[-0.388, -0.037]$
host/donor unseen	2	15	14	-0.462 $[-0.587, -0.333]$	-0.084 $[-0.189, +0.000]$

The true-mask gap is negative in every cohort and replication, with all reported intervals excluding zero. In the strict subset the self-cited mean also has the same sign in all three replications, but the seed-2 interval includes zero. Thus the results support a ranking-level true-region dissociation that is not solely a monotone probability-calibration artifact; they provide weaker evidence for self-citation on the smallest source-clean cohorts. They do not match the arms' training distributions, eliminate every form of memorization or recover a causal decomposition of the network.

P.3 JOINT HOST/DONOR CONSTRUCTION DEPENDENCE

A further exploratory sensitivity uses the graph of *all* 200 ledger edits: photograph identities are vertices and host–donor relationships are edges, regardless of which role a photograph has elsewhere. Copy-moves add self-edges. The graph contains 120 photographs in 30 connected components. Building it before subsetting preserves known dependence paths through non-evaluation rows. The 23 absent donor photographs still use manifest IDs; this grouping does not add pixel validation for them.

We retain the same cohort and row-weighted D_R estimand, then bootstrap occupied components instead of hosts, jointly weighting both arms, image states and source-negative references (10,000 draws; seed 20260913). The specification `analysis_specs/review_20260913_components.md` records this post-inspection analysis; `make final-results` emits the complete graph, cohort assignments and numerical results without model calls.

Table 37: Joint host/donor component sensitivity on the same fitted edit-ID models. Largest is the largest occupied component's evaluation-row count.

cohort	seed	groups	largest	D_{M^*} gap	D_{own} gap
edit-ID original	0	16	7	$-0.469\ [-0.591, -0.379]$	$-0.170\ [-0.288, -0.100]$
edit-ID original	1	20	8	$-0.484\ [-0.560, -0.421]$	$-0.212\ [-0.329, -0.104]$
edit-ID original	2	14	12	$-0.435\ [-0.508, -0.347]$	$-0.201\ [-0.262, -0.089]$
host unseen	0	14	5	$-0.519\ [-0.685, -0.402]$	$-0.182\ [-0.308, -0.108]$
host unseen	1	12	3	$-0.506\ [-0.657, -0.419]$	$-0.194\ [-0.367, -0.081]$
host unseen	2	12	8	$-0.459\ [-0.565, -0.349]$	$-0.203\ [-0.285, -0.075]$
host/donor unseen	0	11	3	$-0.562\ [-0.768, -0.391]$	$-0.176\ [-0.343, -0.044]$
host/donor unseen	1	9	2	$-0.500\ [-0.715, -0.321]$	$-0.200\ [-0.390, -0.040]$
host/donor unseen	2	10	4	$-0.462\ [-0.579, -0.333]$	$-0.084\ [-0.184, +0.000]$

Strict cohorts have only 11/9/10 occupied components, with largest row blocks 3/2/4. Their gold-region intervals are $[-0.768, -0.391]$, $[-0.715, -0.321]$ and $[-0.579, -0.333]$; self-cited intervals are $[-0.343, -0.044]$, $[-0.390, -0.040]$ and $[-0.184, 0.000]$. Thus seed 2 self-citation remains unresolved. Nine components in seed 1 particularly limit resampling precision; even narrow intervals here do not validate a population sampling model. Shared training, omitted identity aliases, and construction choices can cause dependencies outside this graph. These are conditional exploratory summaries of existing models, not source-grouped training replications. The latter would assign whole joint source/donor groups to training and evaluation *before fitting* and would be required for a stronger unseen-source claim.

Minimal current-evidence export. `make final-results` regenerates the model-free final tables and writes `results/final_20260913/manifest.json`. It binds authoritative row files by SHA-256, exposes exact inclusion/exclusion IDs and original split fingerprints, and states each promoted estimand's readout, region instrument and uncertainty unit. The embedded provenance relation suffices to reconstruct strict subsets without image files. It includes the literal FFAA lexicon and annotation-part map, empty-region IDs and side, strict-reader classifications and completion bounds, plus hard/feather missing IDs and shared-floor cohorts. Failed native masks and defined empty FFAA regions are different events. Historical original row-bootstrap and three-replication summaries remain labeled as exposed-cohort analyses, subordinate to the current evidence map (Table 3). Exact execution checks verify computations; they do not establish inferential validity or turn a finite audit into a deployment population.

P.4 Paired feathering and missing outputs

Hard and feathered native outputs share 48 sample IDs, from 45 source images. We pair by ID and use the same source-cluster bootstrap. For normalized differences, both full-restoration margins must meet the same floor: the eligible counts at floors 0.01/0.025/0.05/0.075/0.10 are 48/47/46/44/43. A difference of separately filtered cohort means is not used as a paired estimate. The raw necessity change is $-0.034\ [-0.071, -0.001]$, and the floor-0.05 normalized change is $-0.099\ [-0.205, -0.003]$. Both the input construction and its inferred native mask change, so these estimate robustness of the complete native-output audit, not a seam-only effect holding the mask fixed.

Table 38: **Paired native-mask diagnostics.** Each row states its own eligible sample/source count. Intervals resample sources rather than treating multiple edits of one photograph as independent.

paired quantity	pairs	sources	change [95% interval]
feather − hard: raw necessity	48	45	$-0.034\ [-0.071, -0.001]$
feather − hard: IoU	48	45	$-0.030\ [-0.120, +0.062]$
feather − hard: ΔNS	44	42	$-0.035\ [-0.076, +0.001]$
feather − hard: normalized, floor 0.05	46	43	$-0.099\ [-0.205, -0.003]$
hard mask: necessity − sufficiency	50	47	$-0.179\ [-0.219, -0.140]$

Two feathered masks were unavailable: `copymove_148240` and `copymove_299200`. Their effects are undefined for the realized native pipeline, so we do not impute zero. If a defined output could have been obtained, each raw probability difference would lie in $[-1, 1]$. Assigning those extremes bounds the mean across all 50 requested rows by $[+0.000077, +0.080077]$. Using the known hard effects on the missing IDs bounds the full-cohort feather-minus-hard mean by $[-0.0719, +0.0081]$; this includes zero. These are worst-case bounds on a finite-cohort point estimate, not confidence intervals or proof of population robustness. Normalized ratios are not assumed to lie in $[0, 1]$.

P.5 Context dependence from bidirectional interventions

Let x_{10} insert the manipulated pixels in R into the authentic source, and x_{01} restore R in the full manipulated image. On identical stored native-mask reconstructions, the operational interaction is

$$J_R = \mathrm{Nec}(R) - \mathrm{Suf}(R) = p(x) - p(x_{01}) - p(x_{10}) + p(x_{\mathrm{src}}). \tag{23}$$

The raw estimate is -0.179 $[-0.219, -0.140]$ on 50 rows from 47 source images. Target/source scores, mask IoU/area and necessity are checked against the original audit before pairing; the row file also supplies all normalized floor sensitivities. A negative J_R means that the measured effect of the region is smaller in the full manipulated context than when inserted into the source. This includes possibilities such as cue redundancy, interactions and nonlinear score saturation; it does not identify which internal mechanism produced the contrast. It is the reason that a small normalized necessity cannot be subtracted from one to allocate internal evidence to the rest of the image.