MOST OF THE ABLATION GAIN WAS THE BASELINE'S TRAINING BUDGET: A PLACEBO-CONTROLLED AUDIT OF EVIDENCE-GUIDED DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evidence-guided detectors of AI-generated images fuse a forensic evidence map into an image backbone, and the ablation row "without the evidence branch" is read as the contribution of the map. That row is the sum of three terms: how far the baseline was trained, what the fusion module adds without any map, and what the content of the map adds. We measure each term separately. On a frozen cross-generator test set of 22,999 images and five training seeds, our detector beats a linear probe capped at 100 epochs by 18.35 points of true-positive rate at 5% false-positive rate. Training the probe to its own stopping criterion removes 15.77 of those points; a placebo arm, the same module retrained on all-zero maps, accounts for 1.98 more; the content of the map adds 0.60 (two-level 95% interval $[-0.96, +2.20]$). Against the converged probe the fused detector shows no gain distinguishable from zero on clean images and is 0.033 AUC lower under blur; for the module retrained without a map, 0.029 AUC lower, a probe on the module's own input attributes half of that loss, and nearly all of the corresponding TPR loss, to the module's parameters. The map share is the one quantity that does not move: across four placebos, three training budgets and two resampling schemes the upper end of its macro-AUC interval is at most $+0.0103$, and every reported map-share interval that excludes zero does so on the positive side. Two of five seeds nevertheless read the map, and those are the two seeds that lose most under blur. For a published method, LaRE^2, the same ablation read under three seeds and two checkpoint rules spans -0.016 to $+0.060$ AUC and changes sign for two of three seeds. We define the *ablation resolution* of a protocol and show that a single run cannot separate these terms. Our evidence covers one detector we built and one published method; it shows that the problem exists, not how common it is.

1 INTRODUCTION

An ablation table is a causal claim. A row labelled "without the evidence branch" that scores x points lower is read as a statement that the evidence produced x points. That reading holds under three conditions. The baseline in that row must be as well trained as the model it is compared with. The row must remove the evidence and nothing else. And x must exceed what a rerun of the same configuration would produce.

Machine learning has learned to check the first and third conditions on leaderboards. Tuned baselines and new seeds removed reported gains in reinforcement learning (Henderson et al., 2018; Agarwal et al., 2021), language modelling (Melis et al., 2018), sequence tagging (Reimers & Gurevych, 2017), generative adversarial networks (Lucic et al., 2018), and metric learning (Musgrave et al., 2020). In domain generalization the conclusion depends on a model-selection rule that papers often leave unstated (Gulrajani & Lopez-Paz, 2021). Variation from seeds, data order and hyperparameter search can match the differences that papers report (Bouthillier et al., 2021; Dodge et al., 2019). Models that agree in distribution diverge out of distribution when only the seed changes (D'Amour et al., 2022; Sellam et al., 2022), and the choice of metric can create or remove an effect (Schaeffer et al., 2023). Venues have responded with reproducibility programs (Pineau et al., 2021), meta-reviews of evaluation failures (Liao et al., 2021), a proposed track for refutations (Schaeffer et al.,

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.

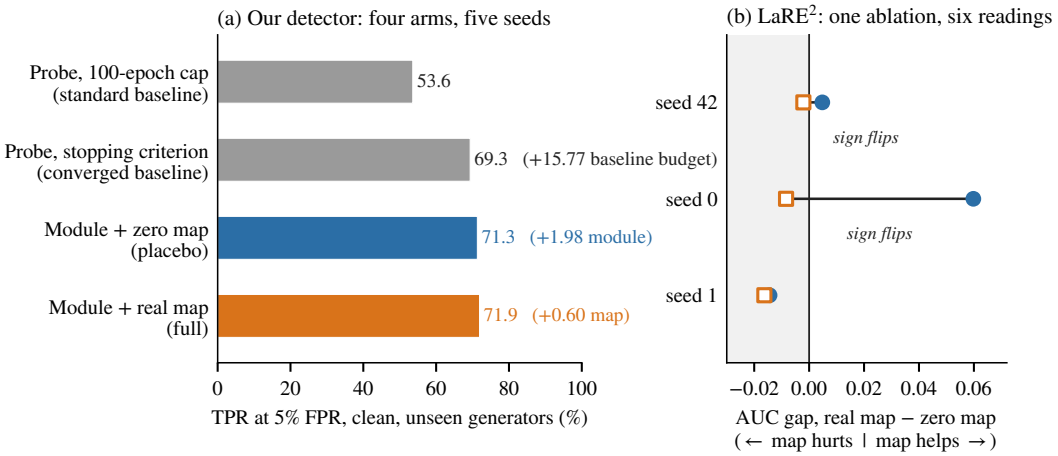


Figure 1: One ablation row, three terms. **(a)** Our detector on clean images from unseen generators, five-seed means. The standard row compares the full model (71.9) with a probe capped at 100 epochs (53.6). Training the probe to its stopping criterion (69.3) accounts for $+15.77$ of the $+18.35$-point gap, the module without a map for $+1.98$ (two-level 95% interval $[-0.78, +4.54]$), and the content of the map for $+0.60$ ($[-0.96, +2.20]$). **(b)** LaRE^2 trained with its released code. Each row is one seed; the two markers are the same ablation read at two stored checkpoints (filled circle: best validation accuracy; open square: latest). The gap spans -0.016 to $+0.060$ AUC and its sign flips for two of three seeds.

2025), and controlled negative findings now change how a subfield reads its results (Yue et al., 2025). Ablation tables have received less of this scrutiny, although Lipton & Steinhardt (2019) list the failure to identify the source of an empirical gain among the recurring flaws of the field. We examine all three conditions in a setting where the second is easy to violate, evidence-guided detection of AI-generated images; the setting is specific, but the three conditions and the instruments that test them are not.

Detectors of generated images must generalize to generators absent from training (Wang et al., 2020a; Chai et al., 2020; Ojha et al., 2023). One line of work gives the detector forensic evidence computed from the image: frequency spectra (Frank et al., 2020), up-sampling traces (Tan et al., 2024), or the error of reconstructing the image with a diffusion model or its autoencoder (Wang et al., 2023; Ricker et al., 2024; Luo et al., 2024). A second line adapts a pretrained vision–language backbone (Radford et al., 2021; Ojha et al., 2023; Liu et al., 2024; Koutlis & Papadopoulos, 2024; Tan et al., 2025; Yan et al., 2025b), and a linear probe on such a backbone is already a strong detector (Ojha et al., 2023). Evidence-guided fusion joins the two: a module reads an evidence map and uses it to refine backbone features, as in LaRE^2 (Luo et al., 2024). When the fused detector beats the backbone, the gain is credited to the evidence. The usual support for that credit is a row without the evidence branch. Such a row removes the map and the parameters that read it, compares the remainder with a baseline trained under some budget, and reports the sum under the name of the map.

We separate the three terms with three instruments. The baseline's budget is measured by training the probe until its own stopping criterion fires instead of an epoch cap. The module is measured by a *placebo ablation*: keep the module, replace the map with an input that carries no information about the image, and retrain with the same seeds. Medicine separates a drug from the act of taking a pill in the same way. The control has relatives in randomization tests of saliency maps (Adebayo et al., 2018), in remove-and-retrain (Hooker et al., 2019), which is why our placebo arm is retrained and not zeroed at test time, in partial-input baselines (Gururangan et al., 2018; Kaushik & Lipton, 2018), and in multimodal analyses of how modalities are used (Hessel & Lee, 2020; Wang et al., 2020b). The content of the map is then the contrast between the full arm and the placebo arm, and its uncertainty is measured by intervals that resample seeds as well as test images (Efron, 1979), read against the smallest effect the protocol can resolve.

The three terms behave differently in our detector (Figure 1a). On a frozen test set of 22,999 images (2,000 real, 20,999 from six generators) and five seeds, the standard row reports $+18.35$ pp TPR at 5% FPR and $+0.038$ AUC over a probe capped at 100 epochs. The converged probe alone recovers $+15.77$ pp and $+0.041$ AUC, so on AUC the fused detector is below the baseline it was meant to improve. The module retrained on zero maps adds $+1.98$ pp ($[-0.78, +4.54]$) and -0.004 AUC ($[-0.008, +0.001]$), and probes on the module's own input do not localize that share (Section 4.1). Under Gaussian blur the fused detector is 0.033 AUC ($[-0.047, -0.016]$) below the converged probe, and half of the module share's AUC loss and nearly all of its TPR loss are attributable to the module's parameters. The map's content adds $+0.60$ pp ($[-0.96, +2.20]$) and $+0.001$ AUC ($[-0.002, +0.003]$). We state the smallness of that share as a bound, not as a null result (Schuirmann, 1987; Lakens, 2017). Across four placebos, three training budgets and two resampling schemes the upper end of every macro-AUC interval is at most $+0.0103$, and the five reported map-share intervals that exclude zero all lie on the positive side. Which intervals exclude zero depends on what is resampled: with the five seeds held fixed the AUC interval is $[+0.0004, +0.0015]$; with seeds resampled it is $[-0.0023, +0.0032]$, and we report both.

A small map share does not mean that the module ignores the map, a distinction familiar from the debate over attention weights (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019). Removing or swapping the map at test time lowers the AUC of two of five full-arm checkpoints by 0.009 and 0.013 and leaves the other three unchanged. Seeds that react more to swapped maps lose more AUC under blur (Spearman $\rho = -0.90$, $n = 5$, exact one-sided $p = 0.042$). We wrote the hypothesis and the test down before evaluating the last two seeds, and both fell on the predicted side. The pattern suggests a shortcut (Geirhos et al., 2020): used by some seeds when present, not needed when absent, harmful when the input degrades.

The same control gives no stable answer for a published method (Figure 1b). We trained LaRE^2 with its released code for three seeds, once with real maps and once with zero maps. The code fixes neither a stopping criterion nor a checkpoint-selection rule, so we read each run at two stored checkpoints. The AUC gap between the two arms ranges from -0.016 to $+0.060$ and changes sign for two of three seeds when the rule changes. A paired bootstrap over images excludes zero for the change in gap in exactly those two seeds, while validation accuracy stays at or above 0.9988 for all six runs. We define the *ablation resolution* of a protocol as the smallest component effect it can separate from variation over seeds, checkpoint rules and test samples. For a single LaRE^2 run it is 0.083 AUC, wider than any map share we measure.

This paper makes four claims, each tied to a check a reader can repeat from the supplementary scores. First, the standard evidence ablation is the sum of a baseline-budget term, a module term and a map term, and the three respond to different experimental choices (Sections 3 and 4.4). Second, in our detector the budget term is the only one distinguishable from zero on clean images; the module term, against the criterion-stopped probe, is not, and it is negative under blur on AUC (Section 4.1). Third, the map term is bounded above by $+0.0103$ AUC under every placebo, training budget and resampling scheme we varied, and the map is read by two of five seeds (Sections 3.2 and 4.3). Fourth, for LaRE^2 a single-seed, single-checkpoint ablation cannot determine even the sign of the map term, because its resolution is coarser than the effect (Section 5).

2 Related Work

Evidence for detecting generated images. Early detectors trained on one generator transfer to others when augmentation is chosen with care (Wang et al., 2020a), and patch-level artifacts explain part of that transfer (Chai et al., 2020). Diffusion models (Rombach et al., 2022) changed the artifacts (Corvi et al., 2023) and motivated reconstruction-based evidence (Wang et al., 2023; Ricker et al., 2024; Luo et al., 2024); Chen et al. (2024) use reconstructions as hard training samples rather than as an input. Benchmarks fix the training generator and test on the rest (Zhu et al., 2023; Yan et al., 2023; Rössler et al., 2019); Yan et al. (2025a) ask whether the task is solved and answer with a harder test set, and we ask, of the same literature, which part of a fused detector produced its improvement.

Controls that separate use from need. Our control follows randomization tests (Adebayo et al., 2018), remove-and-retrain (Hooker et al., 2019) and partial-input baselines (Gururangan et al., 2018; Kaushik & Lipton, 2018); those works test an explanation or a dataset, and we test the attribution made by an ablation row. The closest multimodal analysis (Hessel & Lee, 2020) projects away

cross-modal interaction at test time; the placebo arm removes the content of one input at training time and keeps the parameters.

Variance and claims. Prior work measures how seeds and tuning budgets move headline scores (Henderson et al., 2018; Bouthillier et al., 2021; Dodge et al., 2019; Agarwal et al., 2021), recommends tests for comparing systems (Demšar, 2006; Dror et al., 2018; DeLong et al., 1988), and separates prediction from postdiction (Nosek et al., 2018; Simmons et al., 2011). We move that analysis to the ablation row, where the quantity of interest is a difference of differences and the baseline's budget and the checkpoint rule are two more sources of variation.

3 Three Terms in One Ablation Row

3.1 Setup and the identity

A detector maps an image x to a score $h(x) \in \mathbb{R}$; higher means generated. An evidence-guided detector has three parts: a frozen backbone f, an evidence map $m(x)$ computed from the image, and a trainable fusion module g_θ that reads both, $h_\theta(x) = g_\theta\big(f(x), m(x)\big)$. Our backbone is CLIP ViT-L/14 (Radford et al., 2021); the map is the latent reconstruction error of LaRE^2 (Luo et al., 2024); the module (4.9M trainable parameters) reads the unnormalised global feature after projection together with pre-projection spatial features. Training uses 4,000 images, 2,000 real and 2,000 generated by generators disjoint from the test set, with a 15% validation split, no augmentation, 15 epochs with no stopping criterion, and the checkpoint with the best validation AUC (ties to the earliest epoch); each seed $s \in \{42, 0, 1, 2, 3\}$ changes the initialisation, the data order and the subsample of generated training images. The test set is a frozen manifest of 22,999 images (2,000 real; 20,999 from six generators absent from training), evaluated clean, after JPEG compression at quality 75, and after Gaussian blur. We report two metrics: macro AUC over generators, and the macro average over generators of the true-positive rate at the threshold that yields 5% false positives on real test images (TPR@5%FPR). The macro and pooled definitions differ by 6.5 points on our scores; all TPR values here are macro.

We compare arms trained with the same data, split, optimiser, budget and seed set. B_{100} is the baseline of the standard row: a linear probe on the L2-normalised CLIP embedding, trained with a 100-epoch cap and a validation-loss stopping criterion that never fired. $\mathsf{B}_{\mathrm{stop}}$ is the same probe with the cap raised until the criterion decides the stopping point. F is the full arm; P is the *placebo* arm, the same module trained from scratch with $m(x)$ replaced by an input independent of x. Writing μ_{A} for the expected test metric of arm A over seeds, the gap reported by the standard row splits as

$$\underbrace{\mu_{\mathsf{F}} - \mu_{\mathsf{B}_{100}}}_{\text{standard row}} = \underbrace{\mu_{\mathsf{B}_{\mathrm{stop}}} - \mu_{\mathsf{B}_{100}}}_{\text{budget share } G_{\mathrm{bud}}} + \underbrace{\mu_{\mathsf{P}} - \mu_{\mathsf{B}_{\mathrm{stop}}}}_{\text{module share } G_{\mathrm{mod}}} + \underbrace{\mu_{\mathsf{F}} - \mu_{\mathsf{P}}}_{\text{map share } G_{\mathrm{map}}} . \tag{1}$$

Equation (1) is an identity, not a result; its use is to name what the row estimates. A row that deletes the evidence branch reports the sum and the surrounding text reads it as G_{map}, so the row is biased for the map share by $G_{\mathrm{bud}} + G_{\mathrm{mod}}$, a bias that no number of seeds removes. Only the contrast $\mathsf{F} - \mathsf{P}$ isolates the content of the map. The module share itself compares a probe with a module that reads a different input (bilinear rather than bicubic resizing, no normalisation, spatial features). To keep the second condition of the introduction inside the module term we add two probes on the module's own input: R_{g} on the global feature the module reads, and R_{gs} with the spatial features average-pooled and concatenated, both trained like $\mathsf{B}_{\mathrm{stop}}$. Then $G_{\mathrm{mod}} = (\mu_{\mathsf{R}_{\mathrm{g}}} - \mu_{\mathsf{B}_{\mathrm{stop}}}) + (\mu_{\mathsf{R}_{\mathrm{gs}}} - \mu_{\mathsf{R}_{\mathrm{g}}}) + (\mu_{\mathsf{P}} - \mu_{\mathsf{R}_{\mathrm{gs}}})$, and the last term is the share of the module's parameters. Formal statements are in Appendix A.

3.2 What counts as a placebo

A placebo must satisfy three conditions. (C1) The module, its parameter count and its training budget are unchanged. (C2) The substitute input is independent of the image. (C3) The arm is retrained, so that the module is never evaluated on an input distribution it did not see in training (Hooker et al., 2019), and it is evaluated with the same substitute it was trained on. Zeros satisfy C2 but also remove the variance of the map, so we use four placebos: ZERO; MEAN, the training-set average map; NOISE, Gaussian noise matched to the per-pixel mean and variance of real maps; and SHUFFLE, the real map of another image from the same mini-batch, at training and at test time, which

preserves the marginal distribution of maps and breaks only the pairing with x. Under zero input the map pathway receives no gradient and weight decay closes it during training (Appendix B); the SHUFFLE arm keeps that pathway open and, as Section 4.2 shows, arrives at the same place. Specifications are in Appendix B.

3.3 Intervals that respect both sources of variation

An estimate of any share varies with the seeds that were drawn and with the test images that were drawn. We use a two-level bootstrap (Efron, 1979): each replicate resamples five seed indices with replacement, jointly for all arms, then resamples test images with replacement within each class-by-generator stratum, and recomputes every metric from stored scores, including the 5% threshold; $B = 10{,}000$ replicates, percentile intervals. Resampling images alone answers a narrower question, how these particular checkpoints would score on a fresh test set, and we report that interval beside the two-level one. With five seeds the seed level of a percentile bootstrap is anti-conservative, so for the headline map share we also report the seed-level t interval, $[-0.0035, +0.0054]$ AUC and $[-1.4, +2.6]$ pp, which is wider and under which the headline bound still holds. Where a share is smaller than the protocol can resolve (Section 3.4) we report the upper end of its interval as a bound rather than the point estimate as an effect (Schuirmann, 1987; Lakens, 2017).

3.4 Ablation resolution

A protocol fixes the number of seeds S, the test set, and either one checkpoint-selection rule or none. Let $d_{s,r}$ be the map-share estimate for seed s under rule r, modelled as $d_{s,r} = G_{\text{map}} + a_s + b_r + (ab)_{s,r} + e_T$ with independent zero-mean components of variance $\sigma^2_{\text{seed}}, \sigma^2_{\text{rule}}, \sigma^2_{\text{int}}$ and σ^2_{img}, the last shared by all seeds because they are scored on the same images. For a reader who sees S seeds under one rule the estimator $\bar{d}$ has variance $(\sigma^2_{\text{seed}} + \sigma^2_{\text{int}})/S + \sigma^2_{\text{rule}} + \sigma^2_{\text{img}}$, and the rule term vanishes only when the rule was fixed in advance. We define the *ablation resolution* as the smallest map share detected with power 0.8 at level 0.05,

$$\rho = \left(z_{0.975} + z_{0.8}\right)\sqrt{\tfrac{\sigma^2_{\text{seed}}+\sigma^2_{\text{int}}}{S} + \sigma^2_{\text{rule}} + \sigma^2_{\text{img}}}\,, \tag{2}$$

estimated from stored scores by the method of moments (Appendix A). Only the first term shrinks with S; the image term depends on the test set, and the rule term on whether the rule was pre-specified. With $S \le 5$ a t quantile would enlarge ρ by t_{S-1}/z, so the values we report are lower bounds; a gap below ρ describes a run, not a component.

4 Results on Our Detector

4.1 A converged probe matches the fused detector on clean images and beats it under blur

Claim. Against the criterion-stopped probe, only the budget share of Equation (1) is distinguishable from zero on clean images; the module share is negative under blur on AUC. **Setting.** Five arms, five seeds each, frozen test set; ZERO placebo; all intervals two-level 95%.

The standard row reports $+18.35$ pp and $+0.038$ AUC on clean images (Table 1b); its baseline, the 100-epoch probe, was our own first choice, made and reported before any of the analysis below, not a weak baseline constructed for contrast. The probe trained to its own stopping criterion accounts for $+15.77$ pp and $+0.041$ AUC of that, so on AUC the fused detector is 0.003 below the converged probe and there is no gain to decompose. The module share is $+1.98$ pp ($[-0.78, +4.54]$) and -0.0036 AUC ($[-0.0077, +0.0007]$). Splitting it with the two probes of Section 3.1 does not localize it: R_{g} is $+2.66$ pp ($[+0.14, +5.13]$) above B_{stop}, level with the full arm, but adding the spatial features removes that gain (-2.63 pp, $[-5.05, -0.53]$), leaving $+0.03$ pp for the input representation and, for the module's parameters $\mathsf{P} - \mathsf{R}_{\text{gs}}$, $+1.95$ pp ($[-0.10, +4.34]$) and $+0.0023$ AUC ($[-0.0013, +0.0060]$), both including zero (Table 5). Under blur the converged probe beats the fused detector by 0.0330 AUC ($[-0.0472, -0.0161]$) and 3.21 pp ($[-6.05, -0.31]$), the cap-500 probe by 0.0327 AUC; of the module share, the module's parameters account for -0.0146 AUC ($[-0.0245, -0.0043]$) and -4.03 pp ($[-6.72, -1.38]$), the input representation for the rest (Table 5).

Table 1: Placebo decomposition of the standard ablation row. **(a)** Five-seed means of the arms. **(b)** The three shares of Equation (1) with two-level 95% intervals; the budget-share intervals exclude zero in all six cells and are shown for the clean condition (all six in the supplementary scores).

(a) Arms

Condition	Metric	B_{100}	$\mathsf{B}_{\mathrm{stop}}$	R_{g}	P	F
Clean	AUC	0.9024	0.9431	0.9410	0.9395	0.9405
	TPR	53.58	69.34	72.00	71.32	71.92
Blur	AUC	0.7026	0.7334	0.7210	0.7041	0.7004
	TPR	18.32	23.12	24.95	18.93	19.91

(b) Shares (value; two-level 95% interval beneath)

Condition	Metric	Standard row $\mathsf{F}-\mathsf{B}_{100}$	Budget G_{bud}	Module G_{mod}	Map G_{map}
Clean	AUC	$+0.0381$	$+0.0407$	-0.0036	$+0.0010$
		$[+0.0302, +0.0467]$	$[+0.0366, +0.0450]$	$[-0.0077, +0.0007]$	$[-0.0023, +0.0032]$
	TPR	$+18.35$	$+15.77$	$+1.98$	$+0.60$
		$[+14.95, +21.93]$	$[+13.79, +17.92]$	$[-0.78, +4.54]$	$[-0.96, +2.20]$
JPEG	AUC	$+0.0345$	$+0.0291$	$+0.0023$	$+0.0032$
				$[-0.0067, +0.0115]$	$[+0.0001, +0.0061]$
	TPR	$+6.43$	$+5.47$	-0.55	$+1.51$
				$[-3.76, +2.82]$	$[-0.44, +3.02]$
Blur	AUC	-0.0022	$+0.0308$	-0.0293	-0.0037
		$[-0.0193, +0.0188]$		$[-0.0387, -0.0190]$	$[-0.0119, +0.0044]$
	TPR	$+1.59$	$+4.80$	-4.18	$+0.97$
		$[-1.76, +5.06]$		$[-6.82, -1.38]$	$[-0.43, +2.34]$

The map share, $+0.60$ pp and $+0.0010$ AUC on clean images, is independent of the baseline by construction; that it also stays within one bound under the placebo, budget and resampling choices of Sections 4.2 and 4.4 is the empirical finding. Against the L-BFGS optimum of the same objective the module share keeps its AUC conclusions but not its TPR ones (Appendix B). Per-generator shares and the identity check that the three shares sum to the standard row are in Appendix C.

4.2 The map share stays within the same bound under every placebo and budget

Claim. The upper end of the map-share interval is at most $+0.0103$ AUC under four placebos and three training budgets, and every reported map-share interval that excludes zero is positive. **Setting.** Placebos ZERO, MEAN, NOISE, SHUFFLE, each retrained with the same five seeds, hyperparameters and parameter count; full and ZERO arms retrained to 30 and 60 epochs and read at each budget under the same selection rule.

The four placebos are not distinguishable from one another: the SHUFFLE arm is 0.0037 below the ZERO arm with an interval that includes zero and twice the seed-to-seed spread (Table 2a). Read against SHUFFLE, the only placebo that preserves the marginal distribution of maps, the map share is $+0.0047$ $[+0.0003, +0.0103]$, the one placebo interval that excludes zero; part of that share is the penalty the mismatched input imposes on the SHUFFLE arm itself. The SHUFFLE arm scores the same whether its test maps are swapped or correctly paired (0.9358 in both cases), so a module trained on unpaired maps learns not to read their pairing. Training the full and ZERO arms for twice and four times the budget moves both by at most 0.0014 AUC and leaves the map share between $+0.0010$ and $+0.0019$ with an upper end of at most $+0.0037$ at every budget (Table 2b); one of the six budget intervals (≤ 30, AUC) excludes zero and the ≤ 60 interval includes it again. Across Tables 1 and 2 and the per-generator intervals of Appendix C, the five reported intervals that exclude zero, SHUFFLE, ≤ 30, JPEG, one generator (StarGAN, $+0.0037$ $[+0.0009, +0.0060]$) and the image-only interval, all lie on the positive side. We therefore describe the map's content as a small and probably positive contribution bounded above by $+0.0103$ AUC, rather than as absent; single-generator intervals are wider and are judged in Appendix C. Training details, parameter counts and the weight-decay analysis of the ZERO arm are in Appendix B.

Table 2: Map share on clean images (AUC, five seeds). **(a)** Four placebos; the last row evaluates the SHUFFLE arm on correctly paired maps instead of swapped ones. **(b)** Three training budgets of the full and ZERO arms. Two-level 95% intervals; Bonferroni intervals over the three budgets are 99.17%. All TPR intervals for these rows include zero (Appendix B).

(a) Placebo family Placebo	P AUC	G_{map} [95%]	P − P_{ZERO} [95%]
ZERO	0.9395 ± 0.0046	$+0.0010\ [-0.0023, +0.0032]$	–
MEAN	0.9394	$+0.0010\ [-0.0029, +0.0050]$	–
NOISE	0.9387	$+0.0018\ [-0.0018, +0.0054]$	–
SHUFFLE	0.9358 ± 0.0093	$+0.0047\ [+0.0003, +0.0103]$	$-0.0037\ [-0.0087, +0.0002]$
SHUFFLE, paired at test	0.9358	$+0.0047$	–
(b) Training budget Budget (epochs)	F / P_{ZERO} AUC	G_{map} [95%]	G_{map} [99.17%]
≤ 15 (main)	0.9405 / 0.9395	$+0.0010\ [-0.0023, +0.0032]$	$[-0.0037, +0.0036]$
≤ 30	0.9401 / 0.9382	$+0.0019\ [+0.0006, +0.0032]$	$[+0.0002, +0.0037]$
≤ 60	0.9402 / 0.9388	$+0.0014\ [-0.0003, +0.0031]$	$[-0.0009, +0.0036]$

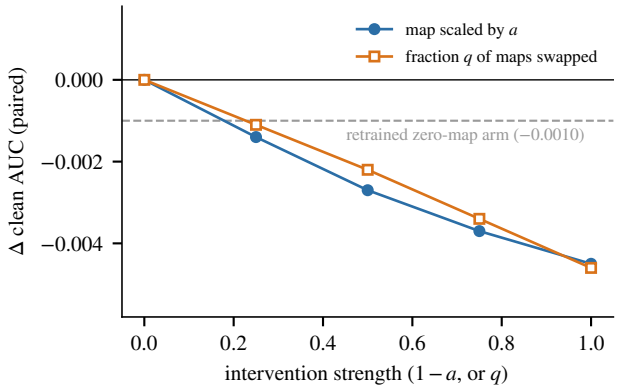


Figure 2: Test-time interventions on the full arm, clean images, five-seed mean of the paired AUC change. Removing the map entirely ($a = 0$) or swapping every map ($q = 1$) costs 0.0045–0.0046 AUC (two-level 95% interval $[-0.0098, -0.0001]$; seed-level t interval $[-0.012, +0.003]$); the loss comes from seeds 0 and 1 (0.009 and 0.013) and the other three seeds move by less than 0.0003 at every level. The dashed line is the retrained ZERO arm.

4.3 TWO OF FIVE SEEDS READ THE MAP, AND THOSE SEEDS LOSE MOST UNDER BLUR

Claim. The full arm's use of the map varies by seed; the seeds that use it lose the most AUC under blur. **Setting.** No retraining. We intervene on the map at test time in the five full-arm checkpoints, scaling it by $a \in \{1, 0.75, 0.5, 0.25, 0\}$ or replacing a fraction $q \in \{0, 0.25, 0.5, 0.75, 1\}$ of maps with those of other images (five permutations), and report paired changes against the same checkpoint's original scores.

Removing the map at test time moves individual scores substantially yet costs only 0.0045 AUC on average, confined to two seeds (Figure 2); the same two seeds react most to swapped maps, and across seeds swap sensitivity tracks the full arm's blur AUC, the measure named in the recorded test (Spearman $\rho = -0.90$, $n = 5$, exact one-sided $p = 0.042$); the blur gap between the full and placebo arms gives the same ranking. With five seeds the smallest attainable exact p is $1/120$, so we make no significance claim and present the result as a prediction that held (Nosek et al., 2018): the hypothesis that a seed which relies on the map inherits its fragility, and the rank test, were recorded before seeds 2 and 3 were evaluated, and both fell on the predicted side (Appendix D). The correlation is not a proxy for training quality: swap sensitivity correlates with clean AUC at only

Table 3: Which conclusions move under which choice. *sign*: the point estimate changes sign; *size*: it changes by more than its interval half-width but keeps its sign; *±0*: the interval's inclusion of zero changes; *stays*: neither; –: not assessed. For the map-share upper-bound row, *stays* means that the interval upper endpoint remains at or below $+0.0103$ AUC.

Quantity	Seeds 3→5	Resampling	Metric	Base. budget	Base. input	Probe stopping	Placebo	Ckpt. rule
Standard row (clean)	stays	–	stays	sign (AUC)	–	–	–	–
Budget share G_{bud}	–	–	stays	–	–	stays	–	–
Module share G_{mod} (clean)	stays	±0 (AUC)	sign	sign	sign	±0 (TPR)	stays	stays
Fused − probe, blur	stays	stays	stays	size ($\times 15$)	stays	±0 (TPR)	–	–
Map share, point estimate	sign	±0	sign (blur)	stays	stays	stays	±0	–
Map share, upper bound	–	stays	stays	stays	stays	stays	stays	–
LaRE^2 gap, sign	sign	–	–	–	–	–	–	sign

-0.50 ($p = 0.225$), and without the best seed the rank correlation over the remaining four is -1.00. The swap protocol and the dated record are in Appendix D; per-seed map shares are in Appendix C.

4.4 What moves and what does not

Claim. Each term of Equation (1) is sensitive to a different experimental choice, and only the bound on the map share survives every choice that Table 3 assesses. **Setting.** No new runs; each entry recomputes a quantity above under one changed choice.

Three statements from our own three-seed analysis did not survive two more seeds: the clean map share changed sign (-0.0004 to $+0.0010$), the placebo arm's advantage in blur-AUC stability fell from eightfold to 1.8-fold, and the blur gap to the capped probe moved from -0.015 to -0.002, whereas against the converged probe it is negative for every subset of seeds (-0.044 for the first three, -0.033 for all five). Table 3 collects these with the choices examined above: the module share changes sign under three of the eight choices, the map share's point estimate under two, and its upper bound under none. Our protocol resolves a clean map share of 0.0045 AUC with five seeds and 0.010 with one (Appendix E), so the map share we measure is below what the protocol can estimate, which is why we report it as a bound. The variance components and the original three-seed subsets are in Appendix E; the full seed-subset analysis is in the supplementary material.

5 Audit of a Published Method

5.1 One ablation, six readings

Claim. For LaRE^2 (Luo et al., 2024), a single-seed, single-checkpoint ablation cannot determine the sign of the map share. **Setting.** Released code and hyperparameters with the local changes of Appendix F, including a two-line switch we added that sets each map to zero; training on a 25,000-image subset of GenImage SD V1.4. Three seeds, each trained once with real maps and once with ZERO maps, for 15 epochs, a budget we set from a pilot run of the real-map arm. We extract the evidence map for all 22,999 frozen test images and evaluate stored checkpoints without retraining. The released code trains without a stopping criterion and does not state which checkpoint the paper reports; in its default configuration its validation and test lists are the same GenImage file, so the stored best-validation and best-test checkpoints coincide in all six runs, and no checkpoint was selected on our frozen test set. We therefore read each run under the two rules the code stores: best validation accuracy (ties to the earliest epoch) and latest.

With the rule held at best-validation, the gap moves from -0.0144 to $+0.0598$ across seeds; changing the rule flips its sign for seeds 42 and 0, and the interval of D excludes zero for exactly those two seeds (Table 4). The largest gap does not come from a better real-map run: it comes from one ZERO-map checkpoint at 0.894 AUC while the other eleven checkpoints lie between 0.93 and 0.96. Validation accuracy is between 0.9988 and 1.0000 for all six runs, so the arms had converged on their training distribution at the budget we set, and the in-distribution signal that the code exposes cannot reveal the instability. The audit uses AUC because the operating point saturates in some runs (Appendix F). Our statement concerns the protocol and not the authors: the code and paper leave the stopping point and the selection rule open, and the conclusion of the ablation depends on both.

Table 4: LaRE^2 audit: AUC gap between the real-map and ZERO-map arms on the frozen test set, per seed and selection rule. D is the best-validation gap minus the latest gap, with a paired bootstrap 95% interval over images (2,000 resamples). The seed-42 real-map run was retrained to 15 epochs to match the other five runs; the 16-epoch pilot gave -0.0015 and $D = +0.0062$ $[+0.0025, +0.0100]$.

Seed	Best-validation	Latest	D	95% interval of D
42	$+0.0048$	-0.0021	$+0.0069$	$[+0.0035, +0.0103]$
0	$+0.0598$	-0.0084	$+0.0682$	$[+0.0627, +0.0738]$
1	-0.0144	-0.0163	$+0.0019$	$[-0.0013, +0.0054]$
Spread over the six readings	0.076 (-0.0163 to $+0.0598$)			

Operating points, the pilot run and the retraining check are in Appendix F; per-checkpoint scores and hashes are in the supplementary material.

5.2 What the protocol can resolve

Decomposing the six readings by the model of Section 3.4 gives $\hat{\sigma}_{\text{seed}} = 0.0092$, $\hat{\sigma}_{\text{rule}} = 0.0101$, $\hat{\sigma}_{\text{int}} = 0.0261$ and $\hat{\sigma}_{\text{img}} = 0.0018$, each with two degrees of freedom and therefore approximate; the seed-by-rule interaction, driven by the seed-0 checkpoint above, is the largest component. The resulting resolution is 0.083 AUC for a single run and rule, 0.053 for the mean of three seeds under one unstated rule, and 0.078 and 0.045 when the rule is fixed in advance (Appendix E). A single reading is compared with the first figure: every one of the six readings, including $+0.0598$, is smaller than what one run of this protocol can separate from a rerun. We make no comparison with the ablation the original paper reports, which uses a protocol and test data we did not reproduce; what we show is narrower: under the released code, changing the seed and the stored checkpoint moves the same ablation across zero.

6 Implications

Each term of an ablation row needs its own instrument, and in our detector each moved with a different choice: the budget share with the rule that stops the baseline, the module share with the baseline's budget, input and stopping, and the LaRE^2 gap with the seed and the checkpoint rule, while the map share stayed within its bound. For evidence-guided detectors this suggests four practices that together cost one additional training run per seed: a baseline trained to its own criterion, a placebo arm, matched inputs for the baseline and the fused model, and a seed set and checkpoint rule fixed before evaluation.

7 Limitations

Our evidence covers one detector we built and one published method on one benchmark family, with five and three seeds respectively, so it establishes that each of the three terms can dominate an ablation row and makes no claim about how often they do; Appendix G pre-specifies the extensions that will widen the sample, at a cost of one training run per seed for each.

8 Conclusion

In our detector, most of the 18.35-point gain that the standard evidence ablation reports came from the baseline's training budget; against the criterion-stopped probe the module's share was not distinguishable from zero on clean images and was negative under blur on AUC; and the map's share was small, probably positive, and bounded above by $+0.0103$ AUC under every placebo, training budget and resampling scheme we varied. For LaRE^2, the same ablation read at three seeds and two checkpoints ranges from -0.016 to $+0.060$ AUC. A single run cannot tell these contributions apart; a converged baseline, a placebo arm, matched inputs and intervals over seeds can, and nothing in them is specific to this domain.

Reproducibility Statement

The test manifest, the seed set and the checkpoint-selection rules were fixed before evaluation, and every experiment added after the first analysis was pre-specified in a dated file (Appendix H). The supplementary material contains the stored per-image scores of every arm, the script that recomputes all tables and intervals from them, and the pre-specification file; Appendix H lists its contents.

Use of Large Language Models

Generative AI tools were used solely to assist with language polishing of the manuscript, including improving the grammar, clarity, readability, and phrasing of author-written text. All AI-assisted revisions were carefully reviewed and edited by the authors to ensure that they accurately reflect the intended scientific meaning. The authors take full responsibility for the final content of this work.

References

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

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, 2021.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European Conference on Computer Vision (ECCV)*, pp. 103–120, 2020.

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

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

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

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

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

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 861–877, 2020.

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

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

Divyansh Kaushik and Zachary C. Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision (ECCV)*, pp. 394–411, 2024.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

Zachary C. Lipton and Jacob Steinhardt. Troubling trends in machine learning scholarship. *ACM Queue*, 17(1):45–77, 2019.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10770–10780, 2024.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are GANs created equal? a large-scale study. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. LaRE^2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17006–17015, 2024.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations (ICLR)*, 2018.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In *European Conference on Computer Vision (ECCV)*, 2020.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018.

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

Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, 22(164):1–20, 2021.

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

Nils Reimers and Iryna Gurevych. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.

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

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.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Rylan Schaeffer, Joshua Kazdan, Yegor Denisov-Blanch, Brando Miranda, Matthias Gerstgrasser, Susan Zhang, Andreas Haupt, Isha Gupta, Elyas Obbad, Jesse Dodge, Jessica Zosa Forde, Francesco Orabona, Sanmi Koyejo, and David L. Donoho. Position: Machine learning conferences should establish a "refutations and critiques" track. In *Advances in Neural Information Processing Systems (NeurIPS), Position Paper Track*, 2025.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. The MultiBERTs: BERT reproductions for robustness analysis. In *International Conference on Learning Representations (ICLR)*, 2022.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

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

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

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020a.

Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020b.

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

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

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

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. DeepfakeBench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70268–70288, 2025b.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

A Formal Statements and Derivations

Observation 1 (what the standard row estimates). Let R be the arm obtained by deleting the evidence branch, both m and g_θ, and training a classifier on the backbone's features under some budget. The gap reported by the row is $\mu_{\mathsf{F}} - \mu_{\mathsf{R}}$. Adding and subtracting $\mu_{\mathsf{B}_{\text{stop}}}$ and μ_{P} gives $\mu_{\mathsf{F}} - \mu_{\mathsf{R}} = (\mu_{\mathsf{B}_{\text{stop}}} - \mu_{\mathsf{R}}) + G_{\text{mod}} + G_{\text{map}}$, so the row read as an estimate of G_{map} has bias $(\mu_{\mathsf{B}_{\text{stop}}} - \mu_{\mathsf{R}}) + G_{\text{mod}}$. The first term is the budget share when R is the capped probe, and it is zero only when R was trained to its criterion. The bias is a property of the contrast, not of its estimator, so no replication removes it. In our detector R and B_{100} are different classifiers on the same backbone: the probe reads a bicubically resized, L2-normalised 768-d embedding, and the module reads a bilinearly resized, unnormalised global feature and pre-projection spatial features. The probes R_{g} and R_{gs} of Section 3.1 realise R on the module's input; on 100 images the two global features have cosine similarity 0.973–1.000 (mean 0.993), almost entirely due to the resizing kernel, with the fp16/fp32 difference contributing cosine > 0.99999.

Proposition 1 (variance of the map-share estimator). Under the model of Section 3.4, the mean of $d_{s,r}$ over S seeds at one rule r drawn from the admissible set has variance $(\sigma^2_{\text{seed}} + \sigma^2_{\text{int}})/S + \sigma^2_{\text{rule}} + \sigma^2_{\text{img}}$; if r is fixed in advance, the rule term is zero. *Argument.* Average the model over s with r and the test set held fixed; b_r and e_T are common to all summands and keep their full variance, while a_s and $(ab)_{s,r}$ average. □

A two-sided level-α test of $G_{\text{map}} = 0$ based on $\bar{d}$ has power $1 - \beta$ at $|G_{\text{map}}| = (z_{1-\alpha/2} + z_{1-\beta})\sqrt{\text{Var}(\bar{d})}$ under a normal approximation, which gives Equation (2). For $S = 1$ the seed and interaction variances cannot be estimated from the protocol itself, which is the formal content of the statement that a single run has no resolution of its own. *Estimators.* For one rule, $\hat{\sigma}^2_{\text{seed}}$ is the sample variance over seeds of d_s and $\hat{\sigma}^2_{\text{img}}$ the variance of the image-only bootstrap of the seed-averaged share. For two rules, the $S \times 2$ table of $d_{s,r}$ is decomposed by the method of moments: $\hat{\sigma}^2_{\text{rule}}$ from the two rule means, $\hat{\sigma}^2_{\text{seed}}$ from the seed means, and $\hat{\sigma}^2_{\text{int}}$ from the residual, which with $S = 3$ has two degrees of freedom and cannot be separated from measurement error. *Assumptions.* Independence of a_s and b_r fails when one rule is unstable only for some seeds, as for seed 0 of LaRE^2; the additive model then assigns that instability to the interaction term. AUC is bounded, so the normal approximation degrades near 1; all AUCs in this paper are below 0.97.

Two-level bootstrap. For $b = 1, \ldots, 10{,}000$: draw five seed indices with replacement, shared across arms; draw images with replacement within each class-by-generator stratum; recompute macro AUC and macro TPR@5%FPR per arm from stored scores, with the 5% threshold recomputed on the resampled real images; store the shares. Intervals are the 2.5th and 97.5th percentiles. Run in image-only mode with the five seeds fixed, the procedure reproduces the interval $[+0.0004, +0.0015]$ previously computed for the clean map share to within 0.00005. Resampling five seeds with replacement underestimates seed variance (the expected bootstrap variance of a mean of n draws is $(n-1)/n$ of the sample variance), so the two-level percentile interval is anti-conservative at the seed level; the seed-level t interval reported in Section 3.3 is the conservative counterpart for the headline share; the other intervals in the paper are two-level percentile intervals. Each analysis runs its own bootstrap with the same random seed but its own set of arms, so a share that appears in two analyses can differ in the last digit.

B Placebo Family, Probes and Training Budgets

Placebo specifications. ZERO replaces the map tensor by zeros in the data loader; MEAN by the per-pixel mean over all training maps, computed once; NOISE by $\mu + \sigma \odot \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$ drawn afresh per sample and μ, σ the per-pixel training statistics; SHUFFLE shifts the maps cyclically within each mini-batch of 16, by an offset drawn from $\{1, \ldots, 15\}$ in training and by one position at evaluation, irrespective of class, so that no image receives its own map and the marginal distribution of maps is unchanged. Evaluation uses the same substitute as training: for SHUFFLE, the evaluation loader is shuffled with five random seeds and rolled within each batch of 16; metrics are averaged over the five permutations inside each bootstrap replicate. Every placebo arm has 4,928,005 trainable parameters, the same as the full arm, and its configuration differs from the ZERO arm's in one line.

The ZERO arm's map pathway. Forward hooks confirm that non-zero maps reach the three map-input layers of every arm. In the ZERO arm the norms of those layers' weights are below 2×10^{-5} of their initial values in all five seeds, whereas the full arm's lie between 0.005 and 0.47; with zero input the gradient of those weights is exactly zero, and restoring the initial weights from the seed and applying weight decay alone for the checkpoint's epochs reproduces the stored norms to four digits (seed 42). Evaluating the ZERO arm on real maps changes individual scores by at most 3.8×10^{-6} and AUC by at most 4.2×10^{-8}.

Table 5: Probes on the module's input and the module's own parameters, five seeds, with two-level 95% intervals beneath the values (one bootstrap, which reproduces the module and map shares of Table 1 exactly). The first three rows sum to the module share.

Difference	Clean AUC	Clean TPR	Blur AUC	Blur TPR
$\mathsf{R}_{\mathrm{g}} - \mathsf{B}_{\mathrm{stop}}$ (resize, normalise)	-0.0021	$+2.66$	-0.0124	$+1.83$
	$[-0.0058, +0.0016]$	$[+0.14, +5.13]$	$[-0.0205, -0.0037]$	$[-0.77, +4.10]$
$\mathsf{R}_{\mathrm{gs}} - \mathsf{R}_{\mathrm{g}}$ (spatial features)	-0.0038	-2.63	-0.0023	-1.98
	$[-0.0070, -0.0001]$	$[-5.05, -0.53]$	$[-0.0111, +0.0069]$	$[-4.15, +0.69]$
$\mathsf{P} - \mathsf{R}_{\mathrm{gs}}$ (module parameters)	$+0.0023$	$+1.95$	-0.0146	-4.03
	$[-0.0013, +0.0060]$	$[-0.10, +4.34]$	$[-0.0245, -0.0043]$	$[-6.72, -1.38]$
$\mathsf{R}_{\mathrm{gs}} - \mathsf{B}_{\mathrm{stop}}$ (input representation)	-0.0059	$+0.03$	-0.0147	-0.15
	$[-0.0098, -0.0019]$	$[-2.81, +2.32]$	$[-0.0230, -0.0071]$	$[-2.11, +2.01]$
Sum $= \mathsf{P} - \mathsf{B}_{\mathrm{stop}}$ (module share)	-0.0036	$+1.98$	-0.0293	-4.18
	$[-0.0077, +0.0007]$	$[-0.78, +4.54]$	$[-0.0387, -0.0190]$	$[-6.82, -1.38]$

Probes. B_{100} stopped at its 100-epoch cap in all five seeds with the criterion (patience 10, minimum improvement 10^{-4} in validation loss, minimum 10 epochs) never firing; its best-loss and best-AUC checkpoints are the same epoch. With the cap raised to 500 the probe reaches 0.9425 ± 0.0018 AUC and 69.11 ± 0.86 pp (JPEG 0.8485 ± 0.0021, 40.58 ± 0.77; blur 0.7331 ± 0.0038, 23.24 ± 0.66), still without the criterion firing. $\mathsf{B}_{\mathrm{stop}}$ (cap 5,000, same criterion, no learning-rate schedule) stopped at epochs 504–622; its checkpoints at epochs 100 and 500 are parameter-identical to the capped runs, so the three are one training trajectory read at three points. R_{g} and R_{gs} use the same criterion and weight decay (10^{-4}) and stopped at epochs 86–194; their training sets are linearly separable, as are $\mathsf{B}_{\mathrm{stop}}$'s. L-BFGS solutions of the same L2-regularised objective converge on all five $\mathsf{B}_{\mathrm{stop}}$ pools, with validation loss 0.0018–0.0048 below the criterion-stopped probe (mean 0.1404 against 0.1436) and test AUC within 0.0007 per seed; on the unnormalised features of R_{g} and R_{gs} the same λ regularises less, nine of ten runs converge (the tenth, R_{gs} seed 2, stops at a gradient of 5.5×10^{-8}), and the solutions are clearly worse than the criterion-stopped probes, so that $\mathsf{P} - \mathsf{R}_{\mathrm{g}}$ becomes $+0.0062$ AUC on clean images and $+0.0108$ under blur. Against the L-BFGS optimum on the $\mathsf{B}_{\mathrm{stop}}$ features the module share keeps its AUC conclusions (-0.0038 $[-0.0080, +0.0004]$ clean, -0.0241 $[-0.0336, -0.0139]$ blur) but not its TPR ones: the clean share moves above zero ($+2.70$ $[+0.003, +5.22]$) and the blur share no longer excludes it (-2.20 $[-4.70, +0.60]$); the fused detector remains below that optimum under blur (-0.0278 $[-0.0421, -0.0110]$ AUC). The split of the module share therefore depends on how the reference probe is stopped, and so does whether its TPR interval includes zero (Table 3); the pre-specified comparison uses the criterion-stopped probes.

Training budget of the fused arms (Table 2b). Full and ZERO arms were retrained to 60 epochs with checkpoints every epoch; the checkpoint selected within the first 15 epochs is parameter-identical to the original in all ten runs, and its evaluation reproduces the main-text intervals to the last digit. Bonferroni 99.17% intervals for TPR are $[-1.67, +2.74]$, $[-0.76, +3.00]$ and $[-0.87, +2.74]$ pp at budgets 15, 30 and 60. TPR map shares for the placebo family are ZERO $+0.60$ $[-0.96, +2.20]$, SHUFFLE $+1.69$ $[-0.47, +4.38]$, and SHUFFLE$-$ZERO -1.09 $[-2.98, +0.54]$; MEAN and NOISE intervals also include zero.

C Per-Generator Shares and Identity Checks

With Bonferroni-corrected 99.17% two-level intervals over the six generators, the clean AUC map share excludes zero for one generator, StarGAN ($+0.0037$ $[+0.0009, +0.0060]$), which shares no

autoencoder with the map extractor, and includes zero for the other five; all six TPR intervals include zero. Single-generator 95% intervals are wider than the macro ones, up to $+0.0150$ AUC at the upper end for DALLE2 (1,000 images). The three shares of Table 1b sum to the standard row exactly in every cell, the five-seed capped-probe means (0.9024, 53.58) agree with the per-seed probe evaluations, and the placebo means agree with the bootstrap's seed-fixed mode. Per-seed map shares ($\mathsf{F} - \mathsf{P}$) are listed in Table 6. The seed-level t interval of Section 3.3 is computed from the

Table 6: Per-seed map share $\mathsf{F} - \mathsf{P}$ for the three conditions and two metrics.

Seed	Clean AUC	Clean TPR	JPEG AUC	JPEG TPR	Blur AUC	Blur TPR
42	-0.005284	-1.94	-0.002490	-0.78	-0.002819	$+1.28$
0	$+0.001560$	$+1.12$	$+0.002481$	$+1.06$	-0.006896	$+0.44$
1	$+0.002422$	$+2.47$	$+0.004016$	$+1.66$	-0.017553	$+0.78$
2	$+0.003486$	$+0.97$	$+0.004396$	$+3.26$	$+0.010497$	$+2.64$
3	$+0.002610$	$+0.39$	$+0.007475$	$+2.33$	-0.001942	-0.26
Mean	$+0.000959$	$+0.60$	$+0.003176$	$+1.51$	-0.003743	$+0.97$

clean rows ($n = 5$, $t_4 = 2.776$).

D Mechanism: Protocols and Records

Swap sensitivity (Section 4.3). The rank statistic uses the 7,980 test images whose maps were cached, each receiving the map of another image drawn over the whole set; a seed's swap sensitivity is the change in its scores under that exchange, and the blur gap is the seed's full-arm blur AUC minus its placebo-arm blur AUC. The interventions of Figure 2 instead exchange maps inside shuffled evaluation batches of 16 (five loader seeds), which matches the training-time shuffle and needs no cached maps; at $q = 0$ this reproduces the stored scores exactly for every image in a full-size batch, and differs by at most 3.1×10^{-4} for the images in the smaller final batches, a batch-size dependence of GPU arithmetic that is an order of magnitude below every effect reported. Paired changes at $a \in \{0.75, 0.5, 0.25, 0\}$ are -0.0014, -0.0027, -0.0037, -0.0045 AUC and at $q \in \{0.25, 0.5, 0.75, 1\}$ are -0.0011, -0.0022, -0.0034, -0.0046; the two-level interval at every point has its upper end just below zero and the seed-level t interval includes zero.

Record of the prediction. A dated message (6 September, 20:30 local time, described in `PREREG.md`) states the hypothesis that a solution which relies on the map inherits the map's fragility, and the test: the rank correlation between swap sensitivity and blur degradation over the seeds not yet evaluated. Seeds 2 and 3 were trained and evaluated afterwards and fell on the predicted side. The recorded test names the full arm's blur AUC as the degradation measure, which the main text reports first; the blur gap between the full and placebo arms gives the same ranking and the same $\rho = -0.90$.

E Seed Subsets and Variance Components

Table 7: Variance components of the per-seed map share and the resulting resolution (Equation (2); $\alpha = 0.05$, $\beta = 0.2$, normal quantiles). Our protocol has one selection rule, so its rule and interaction terms are zero.

Protocol	$\hat{\sigma}_{\text{seed}}$	$\hat{\sigma}_{\text{int}}$	$\hat{\sigma}_{\text{rule}}$	$\hat{\sigma}_{\text{img}}$	$\rho, S = 1$	$S = 3$	$S = 5$
Ours, clean AUC	0.0036	–	–	0.0003	0.0100	0.0058	0.0045
Ours, clean TPR (pp)	1.61	–	–	0.40	4.65	2.83	2.30
Ours, blur TPR (pp)	1.09	–	–	0.50	3.35	2.25	1.95
LaRE^2, clean AUC, rule unstated	0.0092	0.0261	0.0101	0.0018	0.0827	0.0532	0.0450
LaRE^2, clean AUC, rule fixed in advance	0.0092	0.0261	0	0.0018	0.0777	0.0450	0.0350

Over the 31 non-empty subsets of the five seeds, the three statistics of Section 4.4 were recomputed on each subset; `recompute.py` regenerates them all. The subset $\{42, 0, 1\}$ used in our first analysis gives a clean AUC map share of -0.0004, a blur-AUC standard-deviation ratio (full over placebo) of 8.4, and a full-minus-capped-probe blur gap of -0.015 with each seed paired with its own capped probe (-0.014 in the first analysis, which used the seed-42 probe for all seeds); the full set gives $+0.0010$, 1.8 and -0.002. Against the criterion-stopped probe the gap is -0.044 for that subset, -0.033 for all five seeds, and negative for every subset (`E8c_blur_gap_matched.csv` in the supplementary material).

F Audit Details

Code and data. We use the released code with the hyperparameters of its training script. Our local changes, given in full in the supplementary material, are: local paths for the pretrained CLIP weights, which are loaded by name, and for the map annotation file; a fallback in the map-extraction script that assigns an empty class name when a path does not follow GenImage's naming, which leaves the maps unchanged because the class name enters only through a prompt template that we do not enable; a two-line switch that replaces each loaded map by zeros, used for the ZERO arm in training and evaluation; and, for the retrained seed-42 run, a copy of the training script that also saves the epoch-7 weights. Training uses a 25,000-image subset of the GenImage SD V1.4 training split (12,500 real and 12,500 generated; 22,500 for training and 2,500 for validation), obtained from a third-party redistribution and checked against its MD5 checksums. **Checkpoints and rules.** Twelve checkpoints (three seeds, two arms, two rules) plus the 15-epoch retraining of the seed-42 real-map arm. The pilot run of that arm, used to set the 15-epoch budget, stopped at 16 epochs with its latest checkpoint overwritten every epoch; its best-validation checkpoint (epoch 7) is parameter-identical to the retraining's and the first fifteen validation curves coincide, so the retraining reproduces the pilot's trajectory. The stored best-test checkpoints are parameter-identical to the best-validation ones in all six runs because the code's validation and test lists point to the same file; they are therefore not a third rule. **Operating points.** The 95th percentile of real-image scores is 0.9999 (real-map, seed 0, latest), 0.9989 (real-map, seed 42, latest), 0.0000 (ZERO, seed 1, latest) and 0.0000 (real-map, seed 1, best-validation); in the first two runs at least 5% of real images score above 0.99, and in the last two the threshold sits at the floor of the score range, so TPR@5%FPR is uninformative for those runs. **Per-checkpoint AUCs.** Eleven of the twelve checkpoints lie between 0.93 and 0.96 on the frozen test set; the ZERO-map best-validation checkpoint of seed 0 is 0.8938. The twelve AUCs and the checkpoint table are in the supplementary material (Appendix H).

G Pre-Registered Extensions

The following experiments were written into the pre-specification file before submission, with the sentence to be written under each outcome, so that results added during review are not selected after the fact. None of them changes the instruments; each widens the sample the claims are made about.

G.1 LaRE^2, seeds 2 and 3. Same code, budget and rules as Section 5.1. Outcomes: if the five-seed spread under either rule remains above the single-run resolution, the sentence of Section 5.2 is repeated with five seeds; if the spread falls below it, we report that three seeds overstated the instability and give the five-seed value.

G.2 A second published method. A fusion detector whose released code trains an evidence branch and a backbone branch jointly, reproduced to within 1 pp of its reported in-distribution accuracy before any placebo arm is trained; ZERO and SHUFFLE placebos on the evidence input, three seeds, every stored checkpoint read. Outcomes: a map share below that protocol's resolution, or a sign that changes with the rule, is reported as a second instance of Section 5; a map share that the placebo does not recover is reported as the placebo ablation confirming an attribution. Either outcome is reported.

G.3 Fused arms trained to a criterion. F and P with the probe's stopping criterion in place of the 15-epoch cap, five seeds. Outcomes: the module share of Table 1 is recomputed against $\mathsf{B}_{\mathrm{stop}}$; if its sign changes, Table 3 gains a column.

G.4 Ten seeds for our detector. Seeds 4–8 added to the fixed set of five, with the five-seed analysis kept as the primary one and the ten-seed analysis reported beside it. Outcomes: the map-share bound and the resolution at $S = 10$ are reported whatever their values; the two-of-five mechanism of Section 4.3 becomes k-of-ten.

H MANIFEST, PRE-SPECIFICATION AND SUPPLEMENTARY MATERIAL

Manifest. 22,999 images: 2,000 real and 1,000, 5,000, 6,000, 1,999, 1,000 and 6,000 images from six generators; identical evaluation lists for all seeds and arms; batch size 16 for the stored scores of the fused and placebo arms and 64 for those of the probes; TPR@5%FPR is the macro average over generators at the threshold set by the 95th percentile of real scores. **Pre-specification.** Each experiment added after the first analysis (the placebo family, the probes on the module's input, the extended budgets, the test-time interventions, the per-generator analysis, and the criterion-stopped baseline) was committed with its hypothesis, analysis and outcome sentences before its first run; the ZERO and SHUFFLE arms and the swap-sensitivity diagnostic of the first analysis predate the file. The file records every departure from those sentences with its reason: the margin $\delta^\star$ replaced by bounds in the units of the metric, because it depends on the baseline; the per-generator and placebo-family criteria changed from mean differences to intervals, Bonferroni-corrected for the generators; the SHUFFLE arm re-evaluated with the substitute it was trained on, its earlier matched-map reading kept as a separate row; the interventions judged by paired differences within a checkpoint, with maps swapped inside evaluation batches, and the $q = 0$ check restricted to full batches of 16; all six budget intervals reported because the pre-written sentences did not cover a non-monotone result; the LaRE^2 variance decomposed into seed, rule and interaction terms; the best-test column dropped because it equals best-validation; the optional parameter-matched head replaced by the probes on the module's input; the results freeze moved by one day, from 23 to 24 September, before any result of those probes existed, so they were designed and run after the originally scheduled freeze; the recorded test of Section 4.3 naming the full arm's blur AUC, which we report first; and five planned analyses dropped: a second published method (its reproduction could not be attempted for lack of disk space and pretrained weights; Appendix G restores it as an extension), a survey of reporting practice, a corruption-strength sweep, a coverage check, and an interval on a summary-statistics probe. The two-level bootstrap, the seed-subset analysis, the variance components behind the resolution, the diagnosis of the ZERO arm's map pathway and the L-BFGS solutions on the module's input were designed after seeing results and are listed in the file as exploratory; the splits $\mathsf{R}_\mathrm{g} - \mathsf{B}_\mathrm{stop}$ and $\mathsf{R}_\mathrm{gs} - \mathsf{R}_\mathrm{g}$ were pre-specified as reported quantities without outcome sentences, and the L-BFGS reference for B_stop was pre-specified. **Supplementary material.** (i) per-image scores with image identifiers and generator names for every arm, seed, condition and checkpoint used in any table, our detector and LaRE^2 (`scores_bootstrap/`, `scores_lare*/`, `mapmodule/scores/`), the frozen manifest, and the result files behind every table (`mapmodule/analysis/`); (ii) the analysis scripts and `recompute.py`, which runs them on (i) and compares every output with the stored result files; (iii) `PREREG.md`, the pre-specification file, with the hashes and dates of the commits that changed it in `commits.txt`; (iv) `checkpoints.csv`: score file, seed, condition, selection rule and checkpoint hash for every score file; (v) `lare2_local_changes.diff`, our changes to the LaRE^2 code.