PERIODIC GRADIENT-VARIANCE SUPPRESSION ON SVD-ADAPTED RESIDUALS FOR CROSS-GENERATOR DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

As deepfake generation methods diversify rapidly, detectors trained on a fixed set of manipulation techniques increasingly fail to generalize to unseen generation paradigms, undermining their real-world reliability. Parameter-efficient adaptation of frozen vision backbones, such as SVD-decomposed low-rank residuals, enables efficient deepfake detection but offers no mechanism to distinguish generalizable forgery signal from manipulation-specific shortcuts within the few directions it is allowed to adapt. We study whether environment-conditioned gradient disagreement across known manipulation methods can identify which trainable spectral coefficients in a frozen CLIP residual are reliable, and whether this signal differs meaningfully from temporal gradient instability, the criterion used by recent work. Through a controlled, same-seed comparison, we find that suppressing high-disagreement directions fails when applied once, but periodically refreshing the suppression mask matches a Fishr-style cross-environment baseline across three paired seeds. At matched intervention strength, this criterion also outperforms random and magnitude-based suppression controls, supporting an effect beyond simple capacity reduction. A matched FMSD-style temporal control performs consistently below the cross-manipulation approaches in the evaluated seeds, suggesting the environment signal contributes information beyond gradient instability alone. Leave-one-manipulation-out (LOMO) evaluation reveals a substantial transfer asymmetry that is not explained by simple gradient similarity between methods, with one manipulation method's detection collapsing to near-chance performance when held out entirely. On four unseen generators spanning diffusion and GAN-based synthesis, our approach improves over baselines concentrated on the diffusion-based DDIM generator, while the diffusion-transformer case remains difficult for all methods tested.

1 INTRODUCTION

The rapid diversification of deepfake generation techniques, spanning graphics-based manipulation, GANs, and increasingly diffusion-based synthesis, has outpaced individual detectors' ability to keep up. A detector trained on a fixed set of manipulation methods routinely fails against an unseen generation paradigm (Wang et al., 2020; Ojha et al., 2023), undermining its reliability in exactly the open-ended, adversarial setting where detection matters most.

Prior work has pursued several directions toward closing this gap: analyzing spatial or frequency-domain artifacts specific to particular generation processes (Qian et al., 2020; Li et al., 2020a), or exploiting physically-grounded cues, such as illumination, difficult for any generator to replicate exactly. A more recent line adapts large pretrained vision-language models such as CLIP, exploiting their broad visual priors while updating few parameters for efficiency (Yan et al., 2025). Among these, Effort demonstrates that a frozen backbone combined with a low-rank, SVD-decomposed residual preserves pretrained knowledge while learning to detect forgeries.

This efficiency comes with an unaddressed cost: by construction, a parameter-efficient adaptation method restricts learning to a small number of directions, in Effort's case, a single trainable scalar per attention projection. Nothing in this design distinguishes a genuinely generalizable direction from

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.

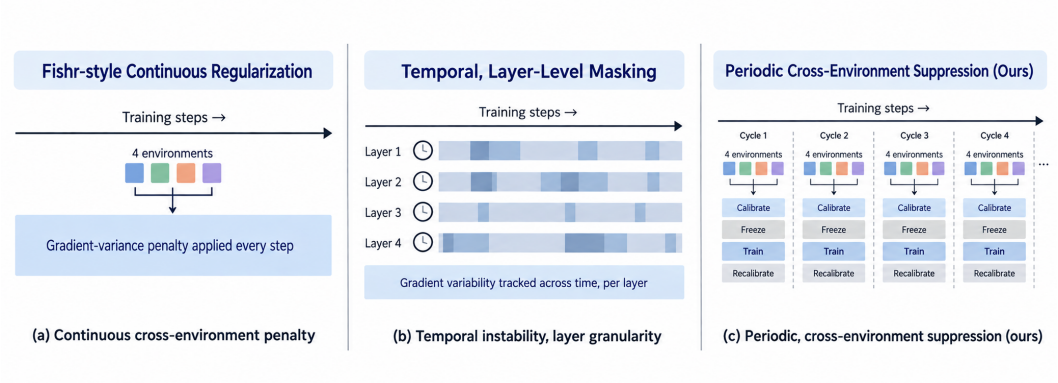


Figure 1: **Comparison of gradient-variance intervention strategies.** (a) Fishr-style continuous regularization applies a cross-environment gradient-variance penalty at every training step. (b) A temporal, layer-level control tracks gradient variability across training time rather than across manipulation methods. (c) Our approach periodically recalibrates a discrete suppression mask using cross-environment gradient disagreement, refreezing it every 150 steps throughout training.

one that has learned a shortcut specific to training-time manipulation methods, precisely the failure that causes generalization to collapse on an unseen generator, with no way for the mechanism itself to detect it.

This relates to invariance-based domain generalization, where cross-environment gradient statistics identify and suppress spurious, domain-specific signal (Ramé et al., 2022a). Recent work has independently explored using gradient statistics to guide which parts of a model to update, measuring instability across training time rather than manipulation methods, at layer rather than coefficient granularity (Zhang et al., 2026). Figure 1 illustrates this distinction. Whether an environment-conditioned criterion on Effort's spectral residual offers a genuinely different signal has not, to our knowledge, been tested.

We show that a one-time suppression mask fails to improve generalization, while periodically recalibrating it throughout training recovers the full benefit of continuous invariance regularization. We confirm this both statistically, across three paired seeds, and geometrically, via a flat-performing weight-space interpolation path. A matched control further shows this benefit is attributable to the cross-manipulation criterion itself, not gradient instability in general.

Our contributions are as follows:

- We show that static suppression of high-disagreement directions fails, while periodically refreshing the same signal matches a continuous invariance regularizer both statistically and, via weight-space interpolation, geometrically.
- We show that this benefit depends specifically on the cross-manipulation-method criterion. A matched control using temporal gradient instability, evaluated on the identical architecture, consistently underperforms.
- We show that the underlying instability measurement is reproducible across training seeds even when the resulting performance rankings are not. This dissociation has implications for how such diagnostics should be interpreted and reported.
- We verify, through random and magnitude-based suppression controls, that this benefit is not explained by capacity reduction alone.
- We report a substantial and only partially-explained transfer asymmetry across manipulation methods, and test, without success, a natural gradient-similarity explanation for it.

2 Related Work

2.1 Parameter-Efficient Deepfake Detection

Deepfake detection has been approached through spatial-domain artifact analysis (Afchar et al., 2018; Li et al., 2020a), frequency-domain decomposition (Qian et al., 2020), and, more recently,

adaptation of large pretrained vision-language models. Examples include MesoNet (Afchar et al., 2018) (mesoscopic properties), F3-Net (Qian et al., 2020) (frequency analysis), SBI (Shiohara & Yamasaki, 2022) (synthetic training pairs), and UCF (Yan et al., 2023a) (common/method-specific feature decoupling) (see Yan et al. (2023b) for a standardized comparison of such detectors).

Frozen-backbone adaptation methods instead exploit CLIP's visual priors (Ojha et al., 2023; Cui et al., 2025) while updating few parameters, following the parameter-efficient adaptation paradigm (Houlsby et al., 2019; Lester et al., 2021; Hu et al., 2022). Layer-norm-only tuning of frozen CLIP has recently been shown to be a strong and even simpler PEFT substrate for this task (Yermakov et al., 2025). MoE-FFD (Kong et al., 2025) introduces mixture-of-experts routing over lightweight adapters while keeping the backbone frozen. Effort instead decomposes each attention matrix via SVD, freezing most singular components and training only a low-rank residual, without a routing mechanism. DiCoME (Kang et al., 2026) similarly builds on frozen CLIP but replaces Effort's uniform decomposition with per-sample projection and evidential fusion. None of these methods identifies which adaptable directions capture generalizable signal versus manipulation-specific shortcuts. Related work suppresses method-specific shortcuts in feature space via probe-SVD (Wang et al., 2026); our intervention instead operates in parameter space, and our static-suppression failure is broadly consistent with their observation that post-hoc remedies can underperform mechanisms coupled to the training trajectory, though their own results show this comparison is not uniform across settings.

2.2 Invariance-Based Domain Generalization

A separate line of work encourages invariance across training domains. Invariant Risk Minimization seeks a representation for which the optimal classifier is identical across environments. This objective can be difficult to optimize reliably in practice (Rosenfeld et al., 2021), and carefully tuned empirical risk minimization remains a strong baseline (Gulrajani & Lopez-Paz, 2021). V-REx (Krueger et al., 2021) penalizes per-domain risk variance directly. Fishr instead matches per-domain loss-gradient variance rather than risk, motivated by the link between gradient covariance and the local Hessian; this signal is the basis for our criterion. Fishr was evaluated within the DomainBed benchmark suite (Gulrajani & Lopez-Paz, 2021); we transplant its signal to a spectral-residual substrate. To our knowledge, this is also the first application of Fishr-style variance-matching to Effort's SVD-decomposed residual; we use it here as a matched-architecture comparison point for the periodic mechanism we propose. Separately, ID3 (Yin et al., 2024) applies invariant-risk-style reasoning to deepfake detection via data augmentation and batch-normalization calibration, not a gradient-based signal. Gradient-based domain generalization has also been applied directly to deepfake detection via perturbed-gradient alignment (Qiu et al., 2025), at whole-model granularity and without a frozen substrate.

Closest to our setting, recent work on forgery-aware layer masking (Zhang et al., 2026) combines frozen-backbone SVD with gradient-statistics-driven masking, but its signal is computed across minibatches over time rather than manipulation methods, at layer rather than coefficient granularity. A related mechanism constructs semantic environments and masks high-variance gradients during language-model reasoning distillation (Cheng et al., 2026), but recomputes the mask every step rather than fixing it, and is not evaluated on visual data. We note that IGA's masking function is structurally identical in form to ours, $M = \exp(-\tau V)$ versus our $m_i = \exp(-\gamma q_i)$; the distinctions we draw are in mask persistence (per-step versus periodic), the normalization and gating applied to the underlying signal, and the substrate (language-model reasoning versus a frozen vision backbone's spectral residual), not in the functional form of the mask itself. Different from these approaches, we study a cross-manipulation-method signal on Effort's spectral residual, intervening periodically at individual-coefficient granularity.

3 Method

3.1 Architecture

Figure 2 illustrates our overall architecture and periodic recalibration pipeline. We build on Effort (Yan et al., 2025), a parameter-efficient adaptation method for a frozen CLIP ViT-L/14 (Radford et al., 2021) backbone. For each attention weight matrix $W \in \mathbb{R}^{d\times d}$ (24 layers, 4 projections: query, key, value, output), Effort applies SVD, $W = U\Sigma V^T$, where $U, V \in \mathbb{R}^{d\times d}$ are orthonormal singular vector

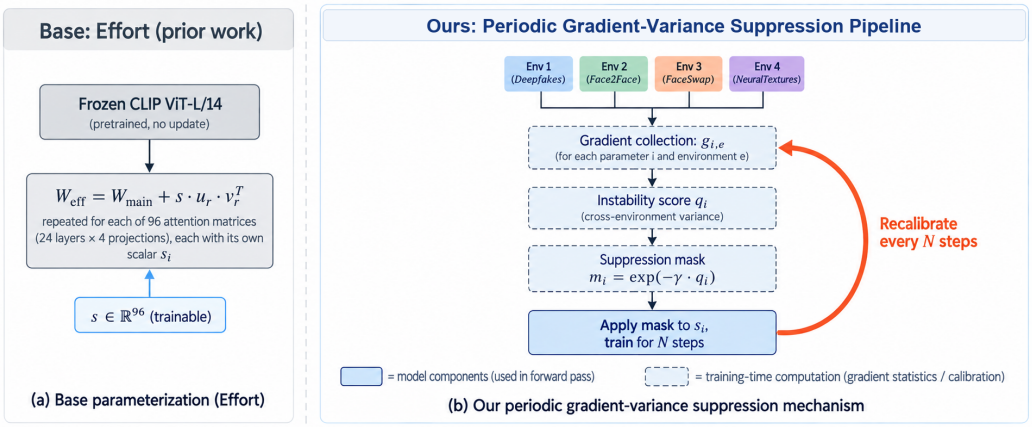


Figure 2: **Overview of architecture and periodic recalibration pipeline.** (a) Base parameterization, adapted from Effort (Yan et al., 2025). (b) Our mechanism: per-environment gradients yield an instability score q_i and mask $m_i = \exp(-\gamma q_i)$, applied and frozen for N steps, then refreshed. Gray: prior work; blue/orange: our contribution. Solid: forward-pass components; dashed: training-time statistics.

matrices and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_d)$ holds singular values in descending order. The decomposition partitions into a principal subspace (top $d-1$ components) and a rank-1 residual:

$$W = \underbrace{U_{[:,:d-1]}\, \Sigma_{[:d-1,:d-1]}\, V^T_{[:,:d-1]}}_{W_{\text{main}}} + \underbrace{u_r\, \sigma_r\, v_r^T}_{W_{\text{residual}}} \tag{1}$$

where u_r, v_r are the final columns of U, V, and σ_r is the smallest singular value. W_{main} is frozen at its pretrained value. Effort's original formulation trains u_r, σ_r, v_r jointly; we instead freeze u_r, v_r at their SVD values and train only σ_r, denoted s. This restricts each matrix's trainable set to a single scalar, yielding 96 total, indexed s_i for $i \in \{1, \ldots, 96\}$. The effective weight at inference and training is $W_{\text{eff}} = W_{\text{main}} + u_r\, s\, v_r^T$. A linear classification head is appended to the pooled output and trained jointly with the 96 scalars via cross-entropy loss. Unless noted, all methods in this paper share this architecture, differing only in how $\{s_i\}$ is trained or intervened upon.

3.2 Environment-Conditioned Instability Criterion

Our central hypothesis is that not all 96 trainable scalars contribute equally generalizable signal: some capture cues that hold across manipulation methods, others capture method-specific shortcuts, and it is the latter we expect to underlie generalization failures. Under gradient starvation (Pezeshki et al., 2021), optimization preferentially captures easily distinguishable features while starving more subtle, generalizable ones; in deepfake detection, manipulation-specific artifacts act as such shortcuts. To distinguish these without unseen data, we exploit a signal available during training: each scalar's gradient consistency across known manipulation methods.

We partition training data into $E = 4$ environments, the canonical FF++ methods (Rössler et al., 2019): Deepfakes, Face2Face, FaceSwap, NeuralTextures. Each environment $e \in \{1, \ldots, E\}$ pairs real images with fakes from that method. For scalar s_i, we compute its gradient per environment, $g_{i,e} = \partial L_e / \partial s_i$, where L_e is the cross-entropy loss on a batch from environment e.

Gradients are accumulated via an exponential moving average (EMA) with decay rate β, $\mu_{i,e}(t) = \beta\, \mu_{i,e}(t-1) + (1-\beta)\, g_{i,e}(t)$, and bias-corrected as $\hat{\mu}_{i,e}(t) = \mu_{i,e}(t)/(1-\beta^t)$. At the end of a calibration window of N steps, we compute the mean $\bar{\mu}_i = \frac{1}{E}\sum_{e=1}^{E} \hat{\mu}_{i,e}$ and population variance $\text{Var}_i = \frac{1}{E}\sum_{e=1}^{E} (\hat{\mu}_{i,e} - \bar{\mu}_i)^2$. To make this scale-invariant across scalars, we normalize by the mean squared gradient, $\tilde{q}_i = \text{Var}_i \,/ (\frac{1}{E}\sum_{e=1}^{E} \hat{\mu}_{i,e}^2 + \epsilon)$, where ϵ prevents division by zero (closely related to the squared coefficient of variation).

A scalar near-zero in every environment would trivially inflate $\tilde{q}_i$ via a near-zero denominator despite carrying no signal. To prevent this, we introduce a minimum-signal gate $a_i = |\bar{\mu}_i| / (|\bar{\mu}_i| + \tau)$, where τ is a threshold constant ($a_i \to 0$ for negligible-signal scalars and $a_i \to 1$ for active scalars), yielding

the final instability score $q_i = a_i \cdot \tilde{q}_i = \frac{|\bar{\mu}_i|}{|\bar{\mu}_i|+\tau} \cdot \tilde{q}_i$. Scalars with high q_i carry meaningful signal but disagree substantially across methods, and are the candidates we suppress.

3.3 Static Suppression and Its Failure

A natural first approach is to compute q_i once, early in training, and apply the resulting mask $m_i = e^{-\gamma q_i}$ for the remainder of training. We refer to this as *static suppression*. We calibrate q_i over an initial 300-step window, freeze the mask, and continue training for a further 500 steps under it.

Evaluated on DF40 (Section 4), static suppression achieves an AUC statistically indistinguishable from an unregularized baseline (0.906 vs. 0.906; Table 1).

We considered three candidate explanations. First, whether suppression was too weak: not the case, as 26 of 96 scalars received a mask value below 0.1, and instability scores showed genuine spread rather than a degenerate concentration. Second, whether the optimizer compensated by growing suppressed scalars' magnitudes to counteract the intervention: we found the opposite, suppressed scalars grew $1.47\times$ versus $1.74\times$ for unsuppressed ones, ruling out compensation. Third, whether the criterion captures a task-irrelevant property such as network depth: we found a depth correlation that weakened substantially after normalizing by residual magnitude, suggesting the raw correlation partly reflects a scale artifact (full analysis in Appendix A.3).

Having ruled out the first two explanations and found only partial support for the third, we hypothesized the failure is attributable to mask *staleness*: an instability snapshot may no longer reflect which directions are unstable as the model adapts. This motivates the mechanism described next.

3.4 Periodic Recalibration

This motivates a mechanism that repeatedly re-estimates and refreshes the suppression mask throughout training rather than fixing it once. We refer to this as periodic recalibration, illustrated in Figure 2(b).

Training proceeds in cycles of $N = 150$ steps. Within each cycle, we reset the EMA statistics $\mu_{i,e}$ (Section 3.2) and accumulate fresh per-environment gradient estimates for all 96 scalars, drawing batches from each of the $E = 4$ environments at every step. At the end of the cycle, we compute q_i as in Section 3.2 and derive a new suppression mask $m_i^{(c)} = \exp(-\gamma\, q_i^{(c)})$, where c indexes the cycle and γ is a fixed suppression-strength hyperparameter ($\gamma = 3.0$ throughout). This mask is held fixed and applied, $s_i^{\text{eff}} = m_i^{(c)} \cdot s_i$, for the following cycle, at whose boundary the process repeats. Over a total training budget of 800 steps, this yields $\lfloor 800/150 \rfloor = 5$ recalibration cycles.

This design differs from static suppression only in re-estimation frequency, and from continuous regularization in that the mask remains fixed and inspectable within each cycle. This allows the mask to be examined directly at any point during training, unlike a continuous penalty with no analogous artifact. This may relate to broader findings that gradient descent implicitly regulates loss sharpness over training (Cohen et al., 2021; Damian et al., 2023), though these results concern full-network loss geometry rather than a low-rank residual specifically, and we do not claim this connection as an established mechanism for our setting.

Like static suppression, periodic recalibration operates entirely during training: the final mask is fixed and applied at inference with no further recomputation. This introduces no additional inference-time cost, a point we return to in Section 6.

4 Experimental Setup

4.1 Data

We train on FaceForensics++ (FF++, c23) (Rössler et al., 2019), where each environment corresponds to a canonical method: Deepfakes, Face2Face, FaceSwap, NeuralTextures (3,000 real and fake images per class from the official split). For evaluation, we construct a 12,000-image manifest from DF40 (Yan et al., 2024) (entire-face-synthesis, Celeb-DF split) with four unseen generators spanning

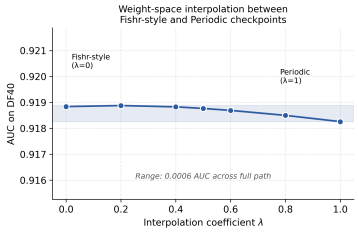


Figure 3: Weight-space interpolation between seed-1 Fishr-style and Periodic checkpoints. AUC remains within 0.0006 across the full path, indicating functionally equivalent solutions.

Table 1: Main results across three seeds, AUC on DF40 (Section 4.1).

Method	Seed 1	Seed 2	Seed 3
Plain Effort	0.9064	—	—
Static suppression	0.9063	—	—
Fishr-style-on-Effort	0.9192	0.9055	0.9168
Periodic Recalibration	0.9191	0.9068	0.9149

Table 2: Paired same-seed Δ (Periodic $-$ Fishr-style).

	Seed 1	Seed 2	Seed 3
Paired Δ	-0.0001	$+0.0013$	-0.0019

diffusion (DiT, DDIM) and GANs (StyleGAN2, StyleGAN3; 1,500 fakes per generator matched with 6,000 Celeb-DF-v2 (Li et al., 2020b) reals).

4.2 Baselines

We compare five methods sharing the architecture in Section 3.1: **Plain Effort** (unregularized baseline), **Fishr-style-on-Effort** (continuous penalty on per-environment gradient variance, detached-gradient approximation (Ramé et al., 2022a)), **FMSD-style control** (reconstruction of recent layer-level temporal masking (Zhang et al., 2026)), **Random suppression** (26 of 96 scalars chosen uniformly at random), and **Magnitude suppression** (suppressing the 26 largest initial-magnitude $|s_i|$ scalars).

4.3 Metrics and Statistical Protocol

Following standard deepfake benchmarking protocols (Yan et al., 2023b), we report AUC and FPR@95%TPR on the 12,000-image manifest, overall and per generator. Seed-to-seed variance exceeding reported effect sizes is a documented pathology in deep learning evaluation (Henderson et al., 2018; Colas et al., 2018); we therefore adopt a paired same-seed protocol throughout to distinguish genuine effects from stochasticity.

5 Results

5.1 Static Suppression Fails, Periodic Recalibration Matches Continuous Regularization

We evaluate whether the environment-conditioned criterion, converted into a suppression mask, improves generalization to unseen generators. Table 1 reports AUC for Plain Effort, static suppression, periodic recalibration, and Fishr-style, each under three seeds, alongside the same-seed paired difference against Fishr-style.

Static suppression achieves an AUC statistically indistinguishable from baseline; periodic recalibration tracks Fishr-style-on-Effort closely across all three seeds. Because raw AUC values alone cannot rule out training stochasticity, we examine the same-seed paired difference directly (Table 2). The paired difference is small at every seed and does not maintain a consistent sign: Periodic Recalibration is marginally below Fishr-style at seeds 1 and 3, and above it at seed 2. A flipping sign with mean near zero (-0.0002) is expected under genuine equivalence rather than an underpowered directional effect; we treat the two methods as empirically equivalent within our precision.

This is corroborated geometrically: since both methods share architecture, data, and initialization, differing only in intervention, their converged parameters can be directly interpolated (Frankle et al., 2020). To our knowledge, this is the first use of weight-space interpolation to test equivalence between a continuous invariance regularizer and a discretized, periodically-applied version of the same signal. We construct $\theta(\lambda) = (1-\lambda)\,\theta_{\text{Fishr-style}} + \lambda\,\theta_{\text{Periodic}}$ for $\lambda \in \{0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0\}$ using seed-1 checkpoints, evaluated on the same DF40 manifest. Figure 3 shows AUC remains within 0.0006 across the full path, smaller than either method's own seed variance. This indicates

Table 3: Same-seed comparison against the temporal, layer-level FMSD-style control. The FMSD-style control was evaluated at two seeds (1 and 3); a third seed was not run.

Method	Seed 1	Seed 3
FMSD-style control	0.9088	0.9081
Fishr-style-on-Effort	0.9192	0.9168
Periodic Recalibration	0.9191	0.9149

Table 4: Mechanistic controls, matched suppression fraction (26/96), seed 1. Granularity-isolation: our criterion at layer-level masking.

Method	AUC	FPR@95%
Plain Effort (no suppression)	0.9064	0.598
Magnitude suppression	0.9063	0.607
Random suppression	0.9098	0.591
Granularity-isolation	0.9183	0.550
Periodic Recalibration	0.9191	0.535

that Fishr-style-on-Effort and Periodic Recalibration converge to functionally equivalent regions of the loss landscape, rather than merely similar-scoring but distinct solutions.

Together, these results establish that a discrete, periodically-refreshed mask recovers the full benefit of continuous regularization, both in aggregate performance and in solution geometry. This is broadly consistent with related findings on forgery-specific shortcut suppression (Wang et al., 2026), though their method operates via post-hoc, inference-time neuron editing rather than a training-time mask, and their own results show a more mixed comparison between remedy types than a single direction of agreement would suggest.

5.2 Cross-Environment Criterion and Mechanistic Controls

We next test whether this benefit is specific to the cross-manipulation-method criterion, or whether any sufficiently dynamic gradient-instability signal would produce comparable results. We construct an FMSD-style control (Section 4.2) applying the identical periodic schedule, architecture, and budget, but with a temporal instability score, tracked across minibatches without partitioning by method, at layer rather than scalar granularity, following recent work on forgery-aware layer masking. This isolates any performance difference to the instability axis and mask granularity, rather than to differences in training procedure.

Table 3 reports same-seed AUC for all three.

At both seeds, both cross-environment methods outperform the control by a consistent margin (Fishr-style: -0.0104/-0.0087; Periodic: -0.0103/-0.0068), exceeding the control's seed-to-seed std (0.0004). This pattern is suggestive of a genuine criterion effect, though with two seeds available, we report it as suggestive rather than confirmed.

The FMSD-style control is our reconstruction from published descriptions. It isolates, within our architecture, whether an environment-conditioned signal offers information beyond a temporal one: it does.

We also test whether comparable gains arise from suppressing any subset of the 96 scalars, regardless of selection, comparing against two controls that suppress the same count (26 of 96) by a fixed magnitude held constant across training.

Random suppression selects the 26 scalars uniformly at random; *magnitude suppression* selects the 26 largest by raw magnitude $|s_i|$ at initialization, reflecting the assumption that the most active parameters most likely encode fragile signal. A *granularity-isolation* control applies our own criterion aggregated to layer-level granularity, isolating whether the benefit depends on fine-grained scalar resolution specifically. Table 4 reports results for all controls alongside Plain Effort and Periodic Recalibration, all at seed 1.

Magnitude suppression provides no benefit over baseline, indicating the largest-magnitude scalars are not preferentially responsible for poor generalization. Random suppression provides a small improvement ($+0.0034$ AUC over Plain Effort), consistent with modest capacity regularization independent of which scalars are chosen. Our criterion substantially outperforms both naive controls at either granularity: the gap over random suppression ($+0.0093$ AUC) is roughly triple random's own gap over baseline, indicating an effect beyond capacity reduction.

The granularity-isolation result clarifies which property, the criterion or its scalar-level granularity, drives the benefit. Granularity-isolation retains our environment-conditioned criterion but aggregates it to layer-level granularity, achieving 0.9183, within 0.0008 of scalar-level Periodic Recalibration

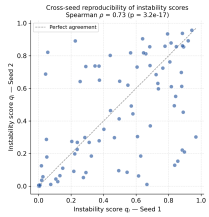


Figure 4: Instability score q_i, seed 1 vs. seed 2, all 96 scalars. Dashed line: perfect agreement. Positively correlated (Spearman $\rho = 0.73$) but with substantial scatter.

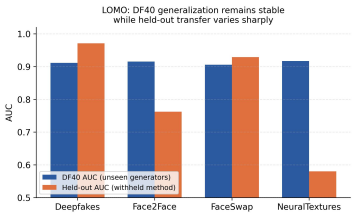


Figure 5: DF40 generalization (blue) remains stable across all four leave-one-out rotations, while held-out transfer performance (orange) varies sharply, collapsing to near-chance when NeuralTextures is withheld.

Table 5: LOMO results. DF40 AUC: generalization under the reduced three-method training set; held-out AUC: performance on the withheld method.

Held-out method	DF40 AUC	Held-out AUC
Deepfakes	0.9116	0.9712
Face2Face	0.9152	0.7621
FaceSwap	0.9059	0.9292
NeuralTextures	0.9172	0.5800

Table 6: Weight averaging ("model soup") of two independently seeded Periodic checkpoints vs. each individually.

Configuration	AUC
Periodic Recalibration, seed 1	0.9191
Periodic Recalibration, seed 2	0.9068
Averaged (seed 1 + seed 2)	0.9158

and well above the FMSD-style control (Table 3). This indicates the benefit is attributable to *which* axis is measured, environment rather than time, not to mask granularity; scalar-level intervention is not necessary for effectiveness, though it remains our preferred, more interpretable choice.

5.3 Diagnostic Reproducibility Does Not Imply Performance Reproducibility

Beyond downstream performance, we examine whether the criterion's flagged scalars are reproducible or stochastic artifacts. We compare the 96 instability scores at the final cycle across two independently seeded runs (Figure 4). The scores correlate significantly (Spearman $\rho = 0.73$, $p = 3.25 \times 10^{-17}$), with 40% and 45% overlap among top-10 and top-20 unstable scalars.

This correlation confirms that strongly disagreeing scalars are reproducibly flagged as a genuine structural property, while moderate-score scalars remain sensitive to initialization and data order. This dissociation—aggregate reproducibility without individual-scalar reliability—is, to our knowledge, not explicitly drawn out in prior work on gradient-variance-based masking criteria, though it echoes broader concerns raised in the machine learning reliability literature (Forde & Paganini, 2019; Pineau et al., 2021). We make it explicit here, supported by both the rank-correlation result above and the flat interpolation path (Section 5.1), and suggest future work reporting similar diagnostics adopt the same distinction.

5.4 Transfer Asymmetry Across Manipulation Methods

To test whether generalization degrades uniformly when a manipulation method is withheld, we conduct leave-one-manipulation-out (LOMO) evaluation across the FaceForensics++ methods. While DF40 generalization remains stable across rotations (0.9059–0.9172), held-out transfer varies drastically: from 0.9712 for Deepfakes down to 0.5800 (near-chance) for NeuralTextures (Table 5, Figure 5), despite NeuralTextures yielding the highest DF40 AUC when included. Pairwise cosine similarity shows NeuralTextures averages 0.912 gradient similarity to remaining methods, indistinguishable from Deepfakes (0.913); gradient dissimilarity cannot explain this failure, leaving it an unexplained negative finding.

5.5 Practical Findings

Weight averaging across seeds improves stability at zero additional training cost. Averaging parameter vectors of two independently seeded Periodic checkpoints yields an AUC of 0.9158 (Table 6), above the weaker seed (0.9068) and nearing the stronger one (0.9191), consistent with weight-averaging findings from independently-trained models (Wortsman et al., 2022; Ramé et al., 2022b; Izmailov et al., 2018; Cha et al., 2021).

Periodic recalibration matches continuous regularization in computational cost. Benchmarking wall-clock time and peak GPU memory shows Periodic Recalibration and Fishr-style-on-Effort incur identical training overhead (8014.8 ms vs. 8005.2 ms/step; Plain Effort: 2049.8 ms) because both accumulate environment gradients at every step (full measurements in Appendix A.7). The practical distinction is interpretability: periodic recalibration produces a fixed, inspectable mask throughout training, whereas continuous regularization provides no such artifact.

6 Discussion and Limitations

Generalization to unseen generators is uneven: performance on StyleGAN2/3 is near-ceiling across methods (including Plain Effort), whereas performance on DiT remains substantially lower for every method tested, indicating diffusion transformers present an unresolved challenge.

The criterion and mechanism combine established components (Effort's SVD residual, Fishr-style gradient variance, CV normalization). Static application fails, whereas periodic recalibration succeeds, outperforms a temporal alternative, and yields a diagnostic reproducible in aggregate but not per-scalar. We consider this characterization our primary contribution.

Our efficiency claim is narrower than initially hypothesized: periodic recalibration matches continuous regularization in cost because both accumulate environment gradients at each step (Section 5.5); the advantage is mask interpretability. Pilot variants exploring warm-starts and scaling to 6,000 steps yielded inconclusive outcomes (Appendix A.6), indicating step scaling alone does not explain performance ceilings.

NeuralTextures' held-out transfer failure in LOMO evaluation (Section 5.4) is not accounted for by gradient dissimilarity, leaving this asymmetry an open question. Finally, our study is evaluated on a single backbone (CLIP ViT-L/14), adaptation architecture (Effort's SVD residual), and domain (deepfake detection); assessing whether these patterns generalize to broader settings remains an open question.

7 Conclusion

We studied whether cross-manipulation-method gradient disagreement identifies which of Effort's trainable spectral coefficients capture generalizable forgery signal versus manipulation shortcuts. Converting this signal into a suppression mask fails statically, but periodic recalibration recovers the full benefit of continuous invariance regularization, confirmed across three paired seeds and via weight-space interpolation. This benefit stems specifically from the cross-environment axis: a matched temporal control underperforms, and naive random/magnitude suppression provide little benefit. The diagnostic is reproducible in aggregate while remaining unreliable per-scalar, a dissociation future work should make explicit. Finally, LOMO evaluation revealed an unexplained transfer asymmetry across manipulation methods. Controlled comparison clarifies what periodic recalibration provides: equivalence to continuous regularization and an inspectable diagnostic, without an efficiency advantage or resolution to diffusion-based generalization failures.

Reproducibility Statement

Our architecture, criterion, and recalibration mechanism are detailed in Section 3; Section 4 specifies datasets. Hyperparameters are reported in Appendix A.2, the paired-seed protocol in Section 4.3, and full per-generator results in Appendix A.5. Anonymized code is provided in the supplementary material.

REFERENCES

Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. MesoNet: A compact facial video forgery detection network. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7, 2018.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. In *Advances in Neural Information Processing Systems*, 2021.

Z. Cheng, W. Dai, and J. Sun. Invariant gradient alignment for robust reasoning distillation. *arXiv preprint arXiv:2606.05025*, 2026.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations (ICLR)*, 2021.

Cédric Colas, Olivier Sigaud, and Pierre-Yves Oudeyer. How many random seeds? statistical power analysis in deep reinforcement learning experiments. *arXiv preprint arXiv:1806.08295*, 2018.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19207–19217, 2025.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations (ICLR)*, 2023.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning*, 2018.

Jessica Zosa Forde and Michela Paganini. The scientific method in the science of machine learning. *arXiv preprint arXiv:1904.10922*, 2019.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, 2020.

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

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

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning (ICML)*, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.

Xiaolu Kang, Zhongyuan Wang, Jikang Cheng, Baojin Huang, Zhanhe Lei, Gang Wu, Qin Zou, and Qian Wang. Divide and conquer: Reliable multi-view evidential learning for deepfake detection. *arXiv preprint arXiv:2606.01885*, 2026.

Chenqi Kong, Anwei Luo, Peijun Bao, Yi Yu, Haoliang Li, Zengwei Zheng, Shiqi Wang, and Alex C. Kot. MoE-FFD: Mixture of experts for generalized and parameter-efficient face forgery detection. *IEEE Transactions on Dependable and Secure Computing*, 2025.

David Krueger, Ethan Caballero, Jörn-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Rémi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (REx). In *International Conference on Machine Learning (ICML)*, pp. 5815–5826, 2021.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face X-Ray for more general face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5001–5010, 2020a.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3207–3216, 2020b.

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

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

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.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision (ECCV)*, pp. 86–103, 2020.

Lingyu Qiu, Ke Jiang, and Xiaoyang Tan. RoGA: Towards generalizable deepfake detection through robust gradient alignment. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Alexandre Ramé, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning (ICML)*, pp. 18347–18377, 2022a.

Alexandre Ramé, Matthieu Kirchmeyer, Thibaud Rahier, Alain Rakotomamonjy, Patrick Gallinari, and Matthieu Cord. Diverse weight averaging for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, 2022b.

Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11, 2019.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18720–18729, 2022.

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

Yihui Wang, Yonghui Yang, Jilong Liu, Fengbin Zhu, Le Wu, and Tat-Seng Chua. Suppressing forgery-specific shortcuts for generalizable deepfake detection. *arXiv preprint arXiv:2606.01843*, 2026.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning*, 2022.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. UCF: Uncovering common features for generalizable deepfake detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22412–22423, 2023a.

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 (Datasets and Benchmarks Track)*, volume 36, pp. 4534–4565, 2023b.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, and Li Yuan. DF40: Toward next-generation deepfake detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 29387–29434, 2024.

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 *PMLR*, pp. 70268–70288, 2025.

Andrii Yermakov, Jan Cech, and Jiri Matas. Unlocking the hidden potential of CLIP in generalizable deepfake detection. *arXiv preprint arXiv:2503.19683*, 2025.

Zixin Yin, Jiakai Wang, Yisong Xiao, Hanqing Zhao, Tianlin Li, Wenbo Zhou, Aishan Liu, and Xianglong Liu. Improving deepfake detection generalization by invariant risk minimization. *IEEE Transactions on Multimedia*, 26:6785–6798, 2024.

Xiang Zhang, Wenliang Weng, Daoyong Fu, Beijing Chen, Ziqiang Li, Ziwen He, and Zhangjie Fu. Generalizable deepfake detection based on forgery-aware layer masking and multi-artifact subspace decomposition. *arXiv preprint arXiv:2601.01041*, 2026.

A Appendix

A.1 AI Usage Statement

Large language model assistance was used in the preparation of this work in the following capacities: aiding and polishing writing for clarity and readability; retrieval and discovery of related work, including verification of methodological claims against primary sources; and research ideation and execution support, including experimental design discussion, debugging of training and evaluation code, and interpretation of results.

A.2 Hyperparameters and Implementation Details

All methods share the architecture described in Section 3.1: CLIP ViT-L/14, with the top 1023 of 1024 singular components frozen per attention matrix and only the rank-1 residual scalar trainable, yielding 96 trainable scalars in total, alongside a linear classification head trained jointly. All methods are optimized with Adam and share the same batch size and image preprocessing (resized to 224×224, normalized using CLIP's standard mean and standard deviation) unless otherwise noted. Table 7 reports the specific hyperparameter values used for our primary experiments (Periodic Recalibration,

Fishr-style-on-Effort, and their associated controls at pilot scale); the full-scale training run described in Section 5 additionally uses a larger per-environment data pool and a proportionally longer training budget, noted separately below the table.

Table 7: Hyperparameters used for pilot-scale experiments (Sections 5.1–5.6, unless otherwise noted).

Hyperparameter	Value
Optimizer	Adam
Learning rate	2×10^{-4}
Batch size	16
Images per environment (train)	3,000
EMA decay (β)	0.9
Minimum-signal threshold (τ)	1×10^{-6}
Suppression strength (γ)	3.0
Calibration window (N)	150 steps
Total training steps	800
Recalibration cycles	5
Random suppression / magnitude suppression, scalars affected	26 of 96

For the full-scale training variant, the per-environment image pool is expanded to the complete available FaceForensics++ training split (approximately 17,000–28,000 images per environment, class-balanced), the total training budget is increased to 6,000 steps, and the calibration window is proportionally lengthened to 500 steps, yielding 12 recalibration cycles; all other hyperparameters are held identical to Table 7.

A.3 Depth-Instability Correlation Analysis

As part of the diagnostic investigation described in Section 3.3 into why static suppression fails to improve generalization, we examined whether the instability score q_i is systematically related to a scalar's depth within the network, that is, which of the 24 transformer layers it belongs to, independent of any forensically meaningful property. A systematic depth relationship of this kind would suggest that our criterion is partly capturing a structural or optimization-related artifact of network depth rather than, or in addition to, genuine cross-manipulation-method disagreement.

We computed the Pearson correlation between each scalar's layer index (1 through 24) and its raw instability score q_i, aggregating across all four attention projections (query, key, value, output) within each layer. This yielded a correlation of $r = 0.54$, indicating a moderate positive relationship between depth and instability under this raw measure. However, this correlation was computed using per-layer mean instability scores, which conflates two distinct sources of variation: genuine differences in cross-environment disagreement, and differences in the raw magnitude of the residual scalar $|s_i|$ itself, which we separately observed to vary systematically with depth. To disentangle these, we recomputed the correlation after normalizing each scalar's instability score by its own residual magnitude, $q_i/|s_i|$, evaluated per-scalar rather than aggregated per-layer. Under this normalized measure, the correlation with depth weakens substantially, to $r = 0.22$.

This weakening indicates that a meaningful portion of the raw depth-instability relationship is attributable to a scale-related artifact, specifically, that residual magnitudes themselves tend to differ systematically with depth, rather than to a genuine forensic pattern in which deeper or shallower layers are inherently more prone to learning manipulation-specific shortcuts. We do not interpret the remaining $r = 0.22$ correlation as ruling out a real depth-related component entirely, but we consider the raw $r = 0.54$ figure alone to be a partially misleading characterization of the underlying pattern, and we report both values here for transparency rather than relying on either in isolation. This analysis was one of three considered explanations for the failure of static suppression discussed in Section 3.3; it provides partial rather than conclusive support for a confounded criterion, and does not by itself account for static suppression's lack of benefit, which we ultimately attribute primarily to mask staleness.

A.4 Static Suppression Diagnostic Chain

Section 3.3 reports that static suppression, a suppression mask computed once early in training and held fixed thereafter, fails to improve AUC over an unregularized baseline. This section reports the

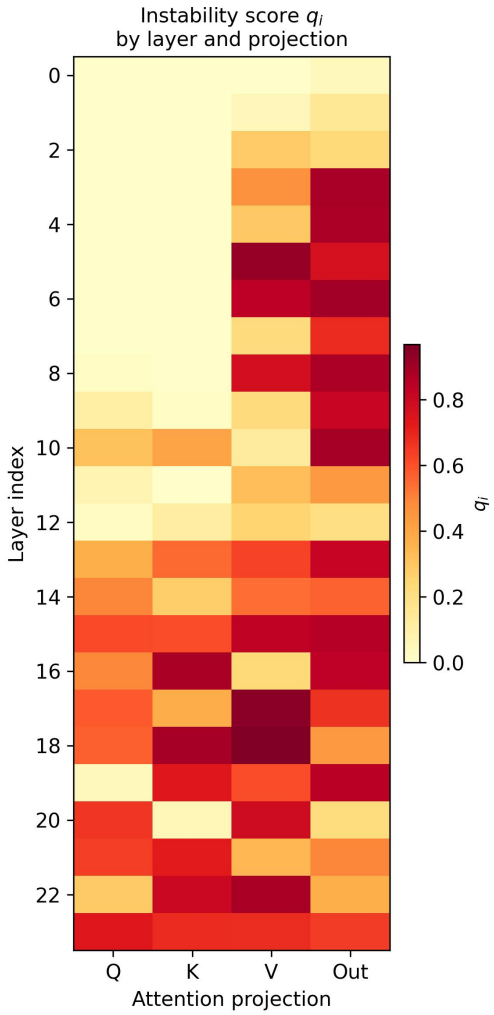


Figure 6: Raw instability score q_i across all 96 scalars, organized by layer and attention projection. Visual counterpart to the depth-correlation analysis above.

full evidence underlying the three candidate explanations considered for this failure, summarized only briefly in the main text.

Insufficient suppression strength. We first considered whether the static mask's suppression was too weak to meaningfully affect the 96 trainable scalars. Of the 96 scalars, 26 received a suppression mask value $m_i < 0.1$, corresponding to a reduction of the affected scalar's effective magnitude to less than one-tenth of its unsuppressed value, a substantial intervention by any reasonable standard. The distribution of computed instability scores $\{q_i\}$ across all 96 scalars showed genuine spread, ranging from near-zero values for scalars with highly consistent cross-environment gradients to values exceeding 0.6 for the most disagreement-prone scalars, rather than a degenerate distribution concentrated near a single value. We consider this evidence sufficient to rule out an under-strength mask as the explanation for static suppression's lack of benefit.

Optimizer compensation. We next considered whether the optimizer compensated for the fixed suppression by disproportionately growing the affected scalars' magnitudes over the remaining training steps, effectively counteracting the intervention. We compared the raw magnitude growth $|s_i^{\text{final}}|/|s_i^{\text{initial}}|$ for the 26 suppressed scalars against the 70 unsuppressed scalars over the final 500 training steps following mask application. Suppressed scalars grew by a factor of $1.47\times$ on average, compared to $1.74\times$ for unsuppressed scalars. This is the opposite of the pattern predicted by an optimizer-compensation explanation, under which suppressed scalars would be expected to grow

more, not less, to counteract their reduced effective contribution. We consider this evidence sufficient to rule out optimizer compensation as the explanation.

A confounded criterion. Finally, we considered whether the instability criterion itself was partly capturing a task-irrelevant structural property, such as network depth, rather than exclusively forensically meaningful cross-manipulation-method disagreement. This hypothesis received partial support, detailed fully in Appendix A.3: raw instability scores correlate with layer depth at $r = 0.54$, though this correlation weakens to $r = 0.22$ after normalizing by residual magnitude, suggesting the raw correlation substantially overstates the extent of any genuine depth confound. We do not consider this evidence conclusive in either direction.

Having found no support for the first two candidate explanations and only partial, inconclusive support for the third, we concluded that static suppression's failure is most plausibly attributable to mask staleness, the possibility that an instability snapshot computed early in training no longer reflects which scalars are unstable as the model continues to adapt. This hypothesis directly motivated periodic recalibration (Section 3.4), whose empirical success (Section 5.1) is consistent with, though does not by itself prove, the staleness explanation.

A.5 Full Per-Seed, Per-Generator Results

This section reports the complete per-generator breakdown underlying every aggregate result presented in Section 5. All values are AUC on the DF40 evaluation manifest described in Section 4.1, broken out by generator (DiT, StyleGAN2, StyleGAN3, DDIM).

Table 8: Full per-seed, per-generator results for all primary methods.

Method	Seed	Overall	DiT	StyleGAN2	StyleGAN3	DDIM
Plain Effort	1	0.9064	—	—	—	—
Static suppression	1	0.9063	—	—	—	—
Fishr-style-on-Effort	1	0.9192	0.757	1.000	1.000	0.945
Fishr-style-on-Effort	2	0.9055	—	—	—	—
Fishr-style-on-Effort	3	0.9168	—	—	—	—
Periodic Recalibration	1	0.9191	0.732	1.000	1.000	0.945
Periodic Recalibration	2	0.9068	—	—	—	—
Periodic Recalibration	3	0.9149	—	—	—	—

Table 9: Full per-seed, per-generator results for controls and additional variants.

Method	Seed	Overall	DiT	StyleGAN2	StyleGAN3	DDIM
FMSD-style control	1	0.9088	—	—	—	—
FMSD-style control	3	0.9081	—	—	—	—
Random suppression	1	0.9098	—	—	—	—
Random suppression	2	0.8948	0.6932	0.9995	0.9991	0.8876
Magnitude suppression	1	0.9063	—	—	—	—
Magnitude suppression	2	0.8948	0.6932	0.9995	0.9991	0.8876
Granularity-isolation	1	0.9183	0.7355	0.9998	0.9997	0.9383
Periodic + Effort-style constraints	1	0.9182	0.7355	0.9998	0.9997	0.9381

Note: Random and Magnitude suppression at seed 2 report identical AUC and per-generator values. This is a genuine empirical result, not a duplication error: both methods, trained independently at this seed, converged to numerically indistinguishable performance. See Section 5.2 for discussion.

A.6 Warm-Start and Full-Scale Training Variants

A warm-start variant, initialized from the released Effort checkpoint rather than a random residual, produced an inconclusive AUC of 0.9119, above the unmodified checkpoint (0.8994) but below our from-scratch result (0.9191). This experiment used a lower learning rate, assuming a converged checkpoint needs only fine-tuning, and did not verify complete fidelity of the loaded residual values via our key-matching diagnostic. Given this unresolved confound, we draw no conclusion from this result and report it alongside a full-scale training variant in Table 10 for completeness.

Table 10: Warm-start and full-scale training variants (reported for completeness, not primary comparisons).

Method	Seed	Overall	DiT	StyleGAN2	StyleGAN3	DDIM
Warm-Start Periodic	1	0.9119	0.7256	0.9988	0.9982	0.9251
Full-Scale Periodic (v1)	1	0.9042	0.7066	0.9998	0.9997	0.9106
Full-Scale Periodic (v2)	1	0.8998	0.6826	0.9997	0.9995	0.9174

A.7 Efficiency Measurement Details

We tested whether periodic computation reduces per-step overhead relative to continuous regularization by measuring wall-clock time and peak GPU memory for all three methods under identical batch size and hardware. Table 11 reports the result.

Table 11: Wall-clock time per step and peak GPU memory, averaged over 20 steps after 5 warmup.

Method	Time (ms)	Memory (MB)
Plain Effort	2049.8	9873
Periodic Recalibration	8014.8	9873
Fishr-style-on-Effort	8005.2	9873

A.8 NT Gradient-Cosine-Similarity Matrix

As discussed in Section 5.4, we tested whether NeuralTextures' anomalously poor held-out transfer performance in the LOMO evaluation could be explained by its gradients being more dissimilar from the other three canonical manipulation methods than those methods are from one another. Table 12 reports the full pairwise cosine similarity matrix between the mean gradient vectors of the 96 trainable scalars, computed under the full four-method training configuration, for each pair of manipulation methods.

Table 12: Pairwise cosine similarity between mean gradient vectors of the 96 trainable scalars, by manipulation method. Values computed under the full four-environment training configuration.

	Deepfakes	Face2Face	FaceSwap	NeuralTextures
Deepfakes	1.0000	0.9120	0.9594	0.8674
Face2Face	0.9120	1.0000	0.9734	0.9511
FaceSwap	0.9594	0.9734	1.0000	0.9175
NeuralTextures	0.8674	0.9511	0.9175	1.0000

NeuralTextures' average cosine similarity to the other three methods is 0.912, and Deepfakes', the best-transferring method in the LOMO evaluation, is 0.913. These averages (and the individual pairwise similarities across all pairs in Table 12) are consistently high and effectively indistinguishable, providing no support for the hypothesis that gradient dissimilarity explains NeuralTextures' held-out transfer failure.

A.9 Code and Supplementary Material Overview

Anonymized code sufficient to reproduce every experiment reported in this work is provided as a supplementary archive. The archive is organized as follows:

- `train_*.py`: training scripts for each method reported in Section 5, one file per method (Plain Effort, static suppression, Periodic Recalibration, Fishr-style-on-Effort, FMSD-style control, random suppression, magnitude suppression, granularity-isolation).
- `eval_df40.py`: the evaluation script used to compute AUC, FPR@95%TPR, and per-generator breakdowns on the DF40 manifest described in Section 4.1, shared across all methods.
- `lomo_*.py`: training and evaluation scripts for the leave-one-manipulation-out experiments reported in Section 5.4.
- `basin_interpolation.py`: the weight-space interpolation script used to produce Figure 3.

- `rank_stability.py`, `nt_gradient_similarity.py`: analysis scripts for the results reported in Sections 5.3 and 5.4, respectively.
- `results/`: JSON files containing the complete numerical results underlying every table in Section 5 and this appendix.

A `README` file accompanying the archive provides further detail on running each script and maps each script to the specific table or figure it produces.

A.10 Extended Statistical Reasoning

The pattern of a sign that flips across seeds while the mean magnitude stays near zero is distinct from what an underpowered but real directional effect would typically produce, which is a consistent sign whose magnitude simply falls below the threshold of statistical confidence at this seed count. Because Fishr-style-on-Effort and Periodic Recalibration share identical architecture, training data, and initialization, differing only in how the suppression signal is applied, their converged solutions occupy the same region of parameter space by construction if the two interventions are functionally equivalent; the flat interpolation path in Figure 3 is the direct empirical test of this prediction, connecting to the broader literature on linear mode connectivity between models sharing an initialization (Frankle et al., 2020; Draxler et al., 2018) and, more generally, the link between flat regions of the loss landscape and generalization (Foret et al., 2021).

A.11 Extended Related Work Differentiation

The distinction from recent work on forgery-aware layer masking is twofold: first, the axis along which instability is measured, our criterion measures disagreement across four labeled manipulation-method environments, while the recent work measures variability across sequential minibatches over the course of training, without reference to which manipulation method produced any given batch; second, the granularity of the resulting mask, our criterion is applied to each of the 96 individual rank-1 residual scalars independently, while the recent work aggregates its signal to the level of an entire transformer layer, coarser by roughly two orders of magnitude. The distinction from the language-model reasoning-distillation mechanism is a difference in temporal persistence: that work recomputes its gradient-conflict mask at every individual optimization step, discarding it immediately after use, whereas our mask is deliberately held fixed across an entire calibration cycle of $N = 150$ steps before being refreshed, a design choice motivated by and validated against the alternative of computing it once and never refreshing it at all (Section 3.3).