RETHINKING VISION FOUNDATION MODELS FOR HIGH-FIDELITY FORGERY DETECTION BY DIAGNOSING OPTIMIZATION COLLAPSE

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision Foundation Models (VFMs) offer a promising route to generalizable deepfake detection, but it remains unclear whether representations learned through general-purpose pretraining retain the subtle, transferable forensic cues needed to detect high-fidelity forgeries. We expose this representational weakness through *Optimization Collapse*, in which linear probes trained on frozen VFM features for high-fidelity forgery detection degrade to chance-level performance even under small adversarial parameter perturbations. We quantify this vulnerability using the perturbation radius at which collapse occurs, termed the *Critical Optimization Radius (COR)*. Theoretical analysis, supported by empirical evidence, shows that the COR is positively associated with the Gradient Signal-to-Noise Ratio (GSNR), a measure linked to generalization, motivating the use of COR as a diagnostic for assessing the strength of transferable forensic cues retained in VFM representations. We validate that higher forgery fidelity exacerbates this optimization vulnerability in both cross-manipulation comparisons and controlled within-manipulation interpolation studies, as reflected in lower COR and poorer detector generalization. Layer-wise analysis further reveals the mechanism underlying *Optimization Collapse*: for high-fidelity forgeries, COR and GSNR progressively decline with depth, indicating that transferable forensic cues are increasingly attenuated as features propagate through the frozen backbone. Reducing the perturbation radius can avoid collapse but does not recover the attenuated forensic cues. To mitigate this collapse by improving GSNR, we propose the *Contrastive Regional Injection Transformer (CoRIT)*, which uses feature differences as a forward-computable proxy for the gradient signal. CoRIT pools token features based on regional gradient-proxy alignment to reduce gradient-proxy variance, injects the pooled features into auxiliary tokens to preserve regional signals, and fuses intermediate and final-layer representations. CoRIT keeps the backbone frozen and optimizes only a lightweight classification head. Experiments show that CoRIT increases both COR and GSNR while achieving the highest average AUC among the compared methods in cross-manipulation and cross-dataset evaluations.

1 INTRODUCTION

The adversarial dynamics between neural synthesis and forensic detection have intensified into a perpetual arms race. The growing diversity of generative models requires forgery detectors that generalize across manipulation techniques. This is challenging because different manipulation processes leave distinct artifact fingerprints (Asnani et al., 2023), and detectors trained on seen artifacts struggle to generalize to unseen techniques. Meanwhile, advances in generative fidelity have made these artifacts increasingly subtle. From GANs (Goodfellow et al., 2014) to diffusion models (Ho et al., 2020) and sophisticated face-swapping methods (Xu et al., 2022), generators produce increasingly realistic content with fewer conspicuous visual artifacts (Frank et al., 2020). This trend calls for detectors capable of extracting subtle yet transferable forensic cues.

Vision Foundation Models (VFMs) have demonstrated broad generalization across downstream visual tasks (Radford et al., 2021). In forgery detection, pioneering work showed that frozen CLIP features support generalizable synthetic-image detection via linear probing (Ojha et al., 2023). Sub-

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.

sequent forgery detection studies have explored parameter-efficient fine-tuning (PEFT) of CLIP (Yan et al., 2025; Cui et al., 2025) and transferable forensic concepts with DINOv3 (Zhou et al., 2026). However, existing studies often overlook a potential mismatch between the general-purpose pretraining objectives of VFMs and forensic analysis of high-fidelity forgeries. The pretraining objectives of VFMs encourage representations of semantically meaningful or visually salient patterns (Radford et al., 2021). In contrast, forensic evidence often resides in perceptually subtle signals, particularly high-frequency anomalies and local texture irregularities (Qian et al., 2020; Guillaro et al., 2023). These pretraining objectives provide no explicit guarantee that frozen VFM features retain generalizable forensic evidence for such subtle traces in high-fidelity forgeries. Instead of costly task-specific fine-tuning, we seek to answer a fundamental mechanistic question: *When and how do frozen VFM representations fail to preserve generalizable forensic evidence for high-fidelity manipulations?*

We observe an optimization failure that exposes this representational weakness while experimenting with Sharpness-Aware Minimization (SAM) (Foret et al., 2021). SAM uses adversarial parameter perturbations to seek flatter regions of the loss landscape and improve generalization, and has also been adopted for training generalizable forgery detectors (Ma et al., 2025). Our unexpected finding is shown in Figure 1(a), where the performance of a linear probe trained with SAM on frozen CLIP-ViT-L/14 features falls to chance level on NeuralTextures (NT) at $\rho = 0.05$, while remaining near-perfect on Face2Face (F2F), DeepFakes (DF), and FaceSwap (FS) under the same training configuration. We term this failure *Optimization Collapse*. Effective learning on NT is restored only after ρ is reduced to 0.01, motivating us to characterize the perturbation radius threshold beyond which collapse occurs, which we term the *Critical Optimization Radius (COR)*. Our key diagnostic insight is that Optimization Collapse is not merely an artifact of an excessively large SAM radius; rather, COR provides an optimization-based stress test for assessing whether frozen VFM representations retain transferable forensic signals.

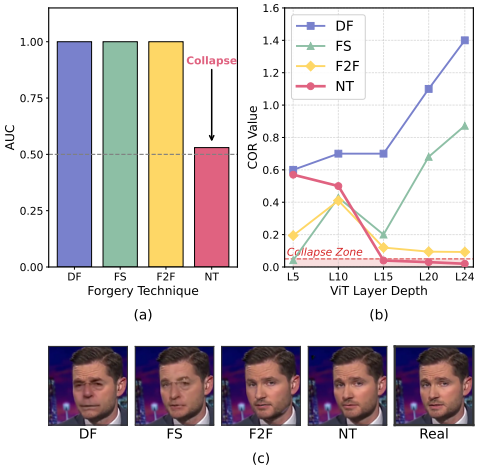


Figure 1: Empirical observation of Optimization Collapse. (a) Detection performance at the SAM radius $\rho = 0.05$. (b) Layer-wise evolution of the COR. The red region denotes the "Collapse Zone" (COR < 0.05). (c) Representative forgery samples illustrating differences in forgery fidelity.

We validate this insight through theoretical and empirical analyses. First, we derive an analytical approximation of COR from a sufficient geometric stability condition that ensures the SAM update remains a valid descent direction. We then establish a theoretical link between COR and the Gradient Signal-to-Noise Ratio (GSNR) (Liu et al., 2020), a statistical measure associated with generalization potential, which shows that the squared COR is proportional to GSNR up to a scaling factor that remains approximately stable during training. Empirical results suggest that the variation of COR is largely driven by GSNR and support the use of COR as an indicator of the generalizable forensic cues encoded in frozen representations. We next examine the relationship between forgery fidelity and generalization through cross-manipulation comparisons and controlled within-manipulation fidelity interpolation, finding that higher-fidelity forgeries exhibit lower COR, smaller semantic displacement, and poorer detector generalization.

Based on this insight, the layer-wise analysis of COR depicted in Figure 1(b) further reveals the mechanism underlying *Optimization Collapse*. For low-fidelity forgeries (DF, FS), which exhibit visually conspicuous artifacts in Figure 1(c), the COR trajectories ascend through the layers, whereas for high-fidelity manipulations (F2F, NT), they rapidly decay. At the final layer, NT's COR falls below 0.05, accounting for its failure in Figure 1(a). This indicates that transferable forensic cues are increasingly attenuated as features propagate through the frozen backbone.

Reducing the SAM perturbation radius can restore stable training but does not enhance the forensic cues retained by the frozen representations. We propose the **Co**ntrastive **R**egional **I**njection **T**ransformer (CoRIT) to counteract their layer-wise attenuation. CoRIT uses feature differences between an original image and its self-blended counterpart (Shiohara & Yamasaki, 2022) as a forward-

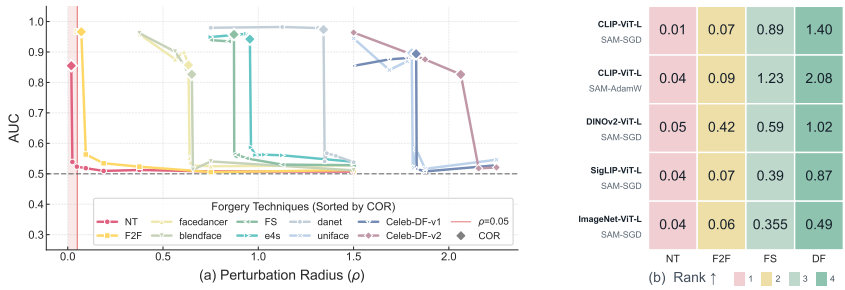


Figure 2: Optimization Collapse observed across different forgery techniques, optimizers, and VFM pretraining strategies. (a) AUC curves of classifiers trained with varying values of ρ across 10 forgery techniques. The red region highlights the setting with $\rho = 0.05$. (b) COR values for different VFMs and optimizers.

computable proxy for the gradient signal, aiming to improve GSNR by preserving regional cues and reducing proxy variance. Based on this proxy, CoRIT groups spatial tokens by regional proxy alignment and pools their features to reduce variance, injects the pooled features into auxiliary tokens through layer-wise residual connections to preserve regional signals, and fuses intermediate representations with deep contextual features. Only the lightweight classification head is trained, without backpropagating through the frozen backbone. Experiments verify that CoRIT increases COR and GSNR alongside improving cross-manipulation and cross-dataset generalization.

Our contributions are as follows. (i) We identify Optimization Collapse and introduce COR as a representation-level diagnostic, with fidelity and layer-wise analyses providing evidence of attenuated transferable forensic cues in frozen VFMs. (ii) We derive an analytical approximation to COR and establish its relationship with GSNR through a state-dependent statistical-geometric scaling factor, connecting optimization stability to gradient signal quality. (iii) We propose CoRIT, a head-only training framework that progressively refines and preserves weak regional cues to mitigate Optimization Collapse, achieving the highest average performance among the compared methods in cross-manipulation and cross-dataset settings.

2 Optimization Collapse: From Optimization Stability to Generalization

In Figure 2(a), a consistent pattern emerges across all 10 forgery techniques: when the SAM perturbation radius ρ exceeds COR, the AUC drops to approximately 0.5, indicating an optimization failure termed *Optimization Collapse*. Figure 2(b) further shows that the ordering of COR values among forgery techniques remains stable regardless of the VFM or optimizer used, with NT consistently exhibiting a small COR.

To characterize this phenomenon, we first recall the optimization mechanism of SAM. Let $\mathbf{w}_t$ denote the parameters of the classification head at step t, and let $\mathcal{L}(\mathbf{w})$ denote the global objective. We write the corresponding gradient and Hessian as $\nabla\mathcal{L}_t = \nabla\mathcal{L}(\mathbf{w}_t)$ and $\mathbf{H}_t = \nabla^2\mathcal{L}(\mathbf{w}_t)$, respectively; $\|\cdot\|$ denotes the Euclidean norm, and $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue. Given a mini-batch $\mathcal{B}_t$, let $\mathbf{g}_t = \nabla\mathcal{L}_{\mathcal{B}_t}(\mathbf{w}_t)$ denote the corresponding stochastic gradient. Standard SAM seeks flat solutions by minimizing the worst-case loss within a Euclidean neighborhood (Foret et al., 2021). Under a first-order approximation, its objective and adversarial perturbation are given by

$$\min_{\mathbf{w}} \max_{\|\boldsymbol{\epsilon}\|_2 \le \rho} \mathcal{L}(\mathbf{w}+\boldsymbol{\epsilon}), \qquad \boldsymbol{\epsilon}_t \approx \rho \frac{\mathbf{g}_t}{\|\mathbf{g}_t\|_2}, \tag{1}$$

where $\rho \ge 0$ is the perturbation radius. Although penalizing sharp minima can improve generalization, an excessively large ρ may perturb the update direction sufficiently to disrupt descent and induce Optimization Collapse. We refer to the corresponding stability threshold as the COR. Specifically, we derive a sufficient stability radius ρ_{critical} from a geometric descent condition and use it as an analytical approximation of the empirical COR. We derive ρ_{critical} as the minimum local stability

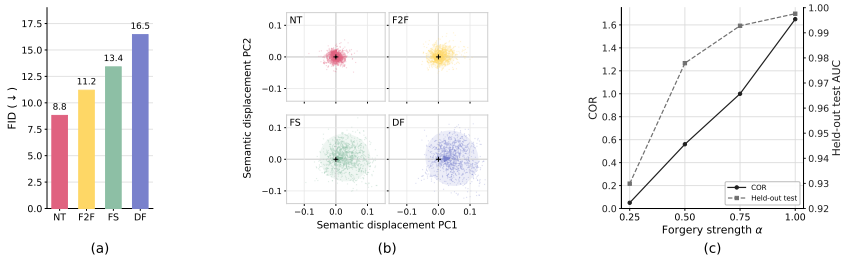


Figure 3: Evidence linking forgery fidelity to generalization. (a) FID across manipulation types. (b) Semantic displacement of forgery cues in a shared CLIP subspace. (c) Within-manipulation trends in COR and held-out AUC across interpolation coefficients α for DeepFakes.

bound along the optimization trajectory:

$$\rho_{\text{critical}} = \min_t \frac{\|\nabla \mathcal{L}_t\|}{\lambda_{\max}(\mathbf{H}_t)}. \tag{2}$$

The *bottleneck step* t^* is the iteration attaining this minimum. At this step, the local curvature imposes the most restrictive bound on the perturbation radius. Evaluating the local bound at t^* yields $\rho_{\text{critical}} = \|\nabla \mathcal{L}_{t^*}\|/\lambda_{\max}(\mathbf{H}_{t^*})$.

The Gradient Signal-to-Noise Ratio (GSNR) (Liu et al., 2020), introduced to quantify generalization capability, measures the strength of the true gradient signal relative to the variance of the stochastic gradient. A smaller GSNR therefore indicates a weaker gradient signal and is theoretically associated with poorer generalization. Following the population-gradient formulation of Liu et al. (2020), we assume that the stochastic gradient is unbiased, i.e., $\mathbb{E}_{\mathcal{B}_t}[\mathbf{g}_t] = \nabla \mathcal{L}_t$. We adopt the following trace-aggregated global variant of GSNR:

$$\text{GSNR}_t = \frac{\|\nabla \mathcal{L}_t\|^2}{\text{Tr}(\text{Cov}[\mathbf{g}_t])}, \tag{3}$$

where $\text{Cov}(\cdot)$ and $\text{Tr}(\cdot)$ denote the covariance operator and the trace, respectively.

We next establish a theoretical connection between COR and GSNR. At the bottleneck step t^*, the squared COR admits the following decomposition into the product of the GSNR and a state-dependent statistical-geometric scaling factor γ_{t^*}:

$$\rho^2_{\text{critical}} = \gamma_{t^*} \cdot \text{GSNR}_{t^*}, \qquad \gamma_{t^*} = \frac{\text{Tr}(\text{Cov}[\mathbf{g}_{t^*}])}{\lambda^2_{\max}(\mathbf{H}_{t^*})}, \tag{4}$$

where γ_{t^*} characterizes the ratio between the statistical scale of stochastic-gradient fluctuations and the geometric scale of local sharpness. Empirical evidence in the appendix indicates that the numerator and denominator of γ_t exhibit closely aligned trajectories, thereby keeping γ_t approximately stable throughout training. Consequently, the dominant source of variation in COR is the GSNR. A vanishing COR may reflect a weak gradient signal and insufficient generalization capability. A complete derivation and detailed interpretation are provided in the appendix.

3 How Forgery Fidelity Shapes Generalization

3.1 Cross-Manipulation Fidelity Evidence

We quantify forgery fidelity using the Fréchet Inception Distance (FID) between real and forged image distributions. Figure 3(a) shows that FID decreases progressively from DF to FS, F2F, and NT, indicating increasing forgery fidelity under this measure. The COR values in Figure 2(b) and the held-out AUCs in Table 1 decrease in the same order. Figure 3(b) further shows that lower FID coincides with smaller paired displacement in a shared CLIP text-semantic subspace, indicating that higher-fidelity forgeries leave subtler semantic cues. These cross-manipulation results reveal a consistent monotonic association: higher forgery fidelity is accompanied by weaker semantic cues and poorer generalization performance.

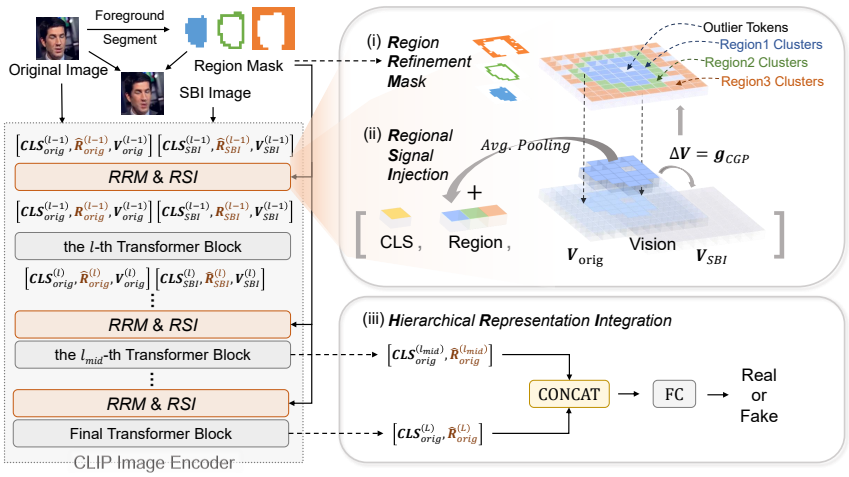


Figure 4: Overview of the proposed **Co**ntrastive **R**egional **I**njection **T**ransformer (CoRIT). Given an original image and its SBI counterpart, both are fed through the frozen CLIP image encoder in parallel. At each transformer layer, CoRIT applies three training-free components: **(i)** The *Region Refinement Mask* clusters visual tokens around region anchors derived from the CGP, filtering out outlier tokens. **(ii)** The *Regional Signal Injection* aggregates the refined tokens via average pooling and injects them into additional Region Tokens through intra-layer residual connections. **(iii)** The *Hierarchical Representation Integration* concatenates the class token and region tokens from an intermediate layer l_{mid} and the final layer L for binary classification.

3.2 Within-Manipulation Fidelity Intervention

To control for manipulation-specific factors in the cross-manipulation evidence, we conduct a within-manipulation intervention on DeepFakes using matched real-forged pairs. Since directly varying the fidelity of a fixed generator is difficult, we use linear interpolation to approximate changes in forgery fidelity: $\mathbf{x}_\alpha = \mathbf{x}_{\text{real}} + \alpha(\mathbf{x}_{\text{forged}} - \mathbf{x}_{\text{real}})$, where $\alpha \in (0, 1]$. A smaller α produces an image closer to its real counterpart, approximating an increase in forgery fidelity. Figure 3(c) shows that both COR and held-out AUC consistently decrease as proxy fidelity increases. This controlled trend provides further evidence that subtler forgery cues impair optimization stability and generalization.

4 Methodology

To counteract this layer-wise representational attenuation, we propose the **Co**ntrastive **R**egional **I**njection **T**ransformer (CoRIT), whose pipeline is illustrated in Figure 4. According to Eq. (3), CoRIT is designed to progressively improve the GSNR across layers by reducing the variance of the gradient proxy and strengthening the gradient-proxy signal. CoRIT enhances both COR and GSNR across layers as shown in Figure 5, thereby mitigating Optimization Collapse.

4.1 Contrastive Gradient Proxy

Maximizing $\text{GSNR}(\mathbf{w})$ as an auxiliary optimization objective is computationally impractical. We therefore shift our analytical focus from the parameter space to the feature space. Following the Self-Blended Image (SBI) paradigm (Shiohara & Yamasaki, 2022), we approximate the gradient using the difference between the representations of an original image and its self-blended counterpart.

$$\mathbf{g}_{\text{CGP}}^{(l)} = \mathbf{V}_{\text{SBI}}^{(l)} - \mathbf{V}_{\text{orig}}^{(l)}. \tag{5}$$

A detailed derivation is provided in the appendix.

4.2 Region Refinement Mask

Because forgery cues are spatially localized, tokens within the same region often exhibit similar CGP directions. We use a Region Refinement Mask to select coherent tokens and apply region-wise average pooling to reduce the variance of regional CGP estimates.

4.2.1 Region Anchor Initialization

Given a feature map consisting of N visual tokens $\mathbf{V}^{(l)} = \{\mathbf{v}_i^{(l)}\}_{i=1}^{N}$, we employ K pre-defined regions, where $\mathcal{N}_k$ denotes the set of token indices in the k-th region. For the k-th region at layer l, we compute a Region Anchor $\mathbf{c}_k^{(l)}$ to capture the dominant gradient direction. Using the CGP of visual tokens $\Delta\mathbf{v}_i^{(l)} = \mathbf{v}_{i,\text{SBI}}^{(l)} - \mathbf{v}_{i,\text{orig}}^{(l)}$, we formulate the anchor as the centroid of the gradients within the region:

$$\mathbf{c}_k^{(l)} = \frac{1}{|\mathcal{N}_k|} \sum_{i \in \mathcal{N}_k} \Delta\mathbf{v}_i^{(l)}, \quad \mathbf{d}_k^{(l)} = \frac{\mathbf{c}_k^{(l)}}{\|\mathbf{c}_k^{(l)}\|_2}, \tag{6}$$

where $|\mathcal{N}_k|$ represents the number of tokens in region k, and $\mathbf{d}_k^{(l)}$ is the normalized directional vector of the region anchor.

4.2.2 Region Refinement and Masking

ViTs may contain outlier tokens (Darcet et al., 2024), which disrupt the coherence of CGP directions within a region. To filter outlier tokens, we refine the region k by clustering tokens around the region anchor $\mathbf{c}_k^{(l)}$. We compute the projection $p_{k,i}^{(l)}$ of each token's proxy gradient onto this anchor direction $p_{k,i}^{(l)} = \langle \Delta\mathbf{v}_i^{(l)}, \mathbf{d}_k^{(l)} \rangle$. The binary region refinement mask is then defined by retaining tokens whose proxy-gradient projections onto $\mathbf{d}_k^{(l)}$ exceed the threshold $\tau\|\mathbf{c}_k^{(l)}\|_2$:

$$\hat{M}_{k,i}^{(l)} = \mathbb{I}(p_{k,i}^{(l)} > \tau \cdot \|\mathbf{c}_k^{(l)}\|_2) \cdot \mathbb{I}(i \in \mathcal{N}_k), \tag{7}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and τ is a hyperparameter controlling the filtering threshold.

4.2.3 Region Average Pooling

We aggregate the selected tokens within each refined region by average pooling to reduce noise. The region token $\mathbf{r}_k^{(l)}$ is computed by averaging the visual tokens $\mathbf{V}^{(l)}$ masked by $\hat{M}^{(l)}$:

$$\mathbf{r}_k^{(l)} = \frac{\sum_i \hat{M}_{k,i}^{(l)} \cdot \mathbf{v}_i^{(l)}}{\sum_i \hat{M}_{k,i}^{(l)} + \epsilon}, \tag{8}$$

where $\mathbf{v}_i^{(l)}$ denotes the i-th visual token in layer l, and $\epsilon = 10^{-6}$ ensures numerical stability.

4.3 Regional Signal Injection

We progressively inject the pooled regional features $\mathbf{r}^{(l)}$ into auxiliary region tokens $\mathbf{R}^{(l)} \in \mathbb{R}^{K \times D}$, allowing regional forensic cues to propagate through the subsequent Transformer layers.

We initialize the region tokens as $\mathbf{R}^{(0)} = \mathbf{0}$. At layer l, the preceding region tokens $\mathbf{R}^{(l-1)}$ are concatenated with the class token $\mathbf{CLS}^{(l-1)}$ and visual tokens $\mathbf{V}^{(l-1)}$. The Transformer block $\mathcal{T}_l$ processes this joint sequence:

$$[\mathbf{CLS}^{(l)}, \hat{\mathbf{R}}^{(l)}, \mathbf{V}^{(l)}] = \mathcal{T}_l\left([\mathbf{CLS}^{(l-1)}, \mathbf{R}^{(l-1)}, \mathbf{V}^{(l-1)}]\right). \tag{9}$$

We update the intermediate region tokens $\hat{\mathbf{R}}_k^{(l)}$ by injecting the regional features $\mathbf{r}_k^{(l)}$ defined in Eq. (8) via an intra-layer residual connection:

$$\mathbf{R}_k^{(l)} = \hat{\mathbf{R}}_k^{(l)} + \mathbf{r}_k^{(l)}. \tag{10}$$

The updated $\mathbf{R}_k^{(l)}$ serves as the input for the subsequent layer $l+1$ in Eq. (9).

Table 1: Cross-manipulation results on FF++ Datasets (frame-level AUC(%); best in **bold**).

Method	Train on remaining three				Avg.
	DF	F2F	FS	NT	
UniFD (Ojha et al., 2023)	95.8	75.0	90.1	59.4	80.1
ForAda (Cui et al., 2025)	97.0	83.3	92.0	64.4	84.2
ForensicConcept (Zhou et al., 2026)	**98.9**	84.5	84.8	62.2	82.6
Ours	98.3	**87.5**	**93.6**	**67.6**	**86.8**

Table 2: Cross-domain performance on classical forensic benchmarks (frame-level AUC (%)). Methods below the midrule share the same SAM optimizer. Best results are in **bold**.

Method	Venues	CDFv1	CDFv2	DFDCP	DFDC	DFD	Avg.
Xception (Rössler et al., 2019)	ICCV'19	77.9	73.7	73.7	70.8	81.6	75.5
F3Net (Qian et al., 2020)	AAAI'20	77.7	73.5	73.5	70.2	79.8	74.9
Face X-ray (Li et al., 2020a)	CVPR'20	70.9	67.9	69.4	63.3	76.6	69.6
SRM (Luo et al., 2021)	CVPR'21	79.3	75.5	74.1	70.0	81.2	76.0
SBI (Shiohara & Yamasaki, 2022)	CVPR'22	-	81.3	79.9	-	77.4	-
UCF (Yan et al., 2023a)	ICCV'23	77.9	75.3	75.9	71.9	80.7	76.3
LSDA (Yan et al., 2024a)	CVPR'24	86.7	83.0	81.5	73.6	88.0	82.6
DiffFake (Sun et al., 2024)	NIPS'24	-	80.5	81.0	-	90.4	-
UDD (Fu et al., 2025)	AAAI'25	-	86.9	85.6	75.8	91.0	-
FreqDebias (Kashiani et al., 2025)	CVPR'25	87.5	83.6	86.8	82.4	74.1	82.9
FIA-USA (Ma et al., 2025)	NIPS'25	90.1	86.7	81.8	-	82.1	-
UniFD (Ojha et al., 2023)	CVPR'23	74.3	74.9	75.5	73.6	86.0	76.9
ForAda (Cui et al., 2025)	CVPR'25	83.7	81.0	81.6	81.6	90.9	83.8
ForensicConcept (Zhou et al., 2026)	ICML'26	85.0	86.9	87.1	83.2	89.7	86.4
Ours	-	**90.9**	**89.1**	**88.8**	**84.5**	**91.3**	**88.9**

4.4 Hierarchical Representation Integration

As shown in Figure 1(b), intermediate layers retain forensic cues that may be attenuated in the final layer. We therefore combine the intermediate-layer forensic representation with the final-layer contextual representation. We define the representation at layer l by concatenating the class token and the region tokens: $\mathbf{F}^{(l)} = [\mathbf{CLS}^{(l)}, \mathbf{R}^{(l)}]$. The output representation is then constructed by concatenating the representations from a predefined intermediate layer l_{mid} and the final layer L:

$$\mathbf{F}_{\text{out}} = \text{Concat}\left(\mathbf{F}^{(l_{\text{mid}})}, \mathbf{F}^{(L)}\right). \tag{11}$$

By preserving forgery representations with high COR, $\mathbf{F}_{\text{out}}$ provides a robust input for final classification and downstream tasks.

5 Experiments

5.1 Datasets and Evaluation Protocols

We assess CoRIT on two complementary benchmark groups. The classical forensic benchmarks comprise FaceForensics++ (FF++) (Rössler et al., 2019), the Deepfake Detection Challenge (DFDC) (benpflaum et al., 2019), its preview set DFDCP (Dolhansky et al., 2019), DeepFake Detection (DFD) (Nick Dufour, 2019), and Celeb-DF (Li et al., 2020b). The appendix provides further details on the emerging-forgery benchmark (Yan et al., 2024b). We evaluate both cross-dataset and cross-manipulation generalization.

5.1.1 Implementation Details.

We use the CLIP-ViT-L/14 image encoder as a frozen backbone and optimize only the classification head with SAM ($\rho = 0.07$), using a batch size of 20 and an initial learning rate of 10^{-3}. Following Shiohara & Yamasaki (2022), we align faces using dlib's 81 landmarks. For RRM, we set $\tau = 0.1$ and $K = 3$, corresponding to the foreground, its boundary, and the background. For HRI, we set $l_{\text{mid}} = 18$. All experiments are conducted on a single NVIDIA V100 GPU.

5.2 Generalization Evaluation

5.2.1 Cross-Manipulation Evaluation

To examine whether VFM representations retain transferable forensic cues for high-fidelity forgeries, we evaluate three VFMs-based comparison methods on FF++ under the leave-one-manipulation-out protocol in Table 1. CoRIT achieves the best average AUC, exceeding UniFD by 6.7 points, CLIP-adapted ForAda by 2.6 points, and DINOv3-based ForensicConcept by 4.2 points, demonstrating its ability to capture transferable forensic cues. For all four methods, held-out AUC consistently follows DF $>$ FS $>$ F2F $>$ NT, matching the COR ranking in Figure 2(b) and the FID ranking in Figure 3(a). This consistent pattern shows that as forgery fidelity increases, both COR and held-out AUC decline, indicating that VFM representations do not reliably preserve generalizable forensic cues for high-fidelity forgeries. Moreover, compared with baseline UniFD, CoRIT gains most on low-COR manipulations, with improvements of 12.5 points on F2F and 8.2 points on NT. This concentration supports COR as a diagnostic of generalization and shows that CoRIT mitigates the high-fidelity representational bottleneck.

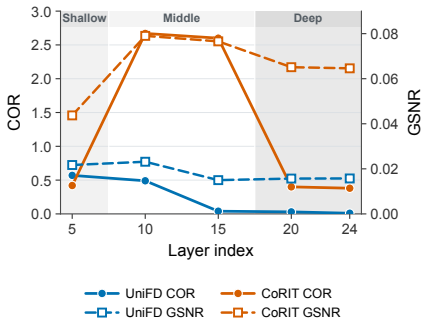


Figure 5: Layer-wise COR and GSNR curves follow similar trends.

5.2.2 Cross-Dataset Evaluation on Classical Forensic Benchmarks

We next assess whether these gains extend to broader cross-dataset shifts. As shown in Table 2, CoRIT achieves the highest average AUC of 88.9%. Its 12.0-point improvement over the matched frozen-CLIP baseline UniFD isolates the benefit of our design, while its 2.5-point advantage over the strongest prior method, DINOv3-based ForensicConcept, shows that the gain is not limited to comparisons within CLIP-based methods. Moreover, freezing the backbone reduces training memory by 90% relative to ForAda, from 21,224 MB to 2,130 MB. Results on emerging forgery techniques from DF40 (Yan et al., 2024b) are provided in the appendix.

5.3 Theoretical Verification and Visualization

We compare the layer-wise COR and GSNR of the baseline UniFD and the proposed CoRIT on NeuralTextures (NT), a challenging dataset characterized by high-fidelity manipulation type. As shown in Figure 5, UniFD's COR rapidly falls below 0.05 by layer 15 and reaches 0.01 at layer 24, indicating Optimization Collapse. In contrast, CoRIT increases COR between layers 10 and 15 and retains a final-layer value of 0.38, remaining well above the theoretical Collapse Zone (COR $<$ 0.05). CoRIT also consistently achieves higher GSNR, with trends aligned with COR, demonstrating its ability to preserve representations of high-fidelity forgery artifacts. Complementary loss-landscape visualizations are provided in the appendix.

5.3.1 Ablation Study of Proposed Components

As shown in Table 3, we progressively validate the contribution of Regional Signal Injection (RSI), Region Refinement Mask (RRM), and Hierarchical Representation Integration (HRI). Compared to the baseline, the introduction of RSI yields a substantial initial gain of 5.2 points. RRM brings an additional 2.9 points by filtering tokens with incoherent CGP. HRI contributes a further 2.1 points, resulting in a total improvement of 10.2 points over the baseline. When applied alone, HRI also improves the baseline by 3.7 points. Each component plays a distinct and complementary role in improving forensic generalization.

Table 3: Ablation study on proposed components (Frame-Level AUC (%)). Best results are in **bold**.

RSI	RRM	HRI	CDFv2	DFDC	facedancer	fsgan	Avg.
-	-	-	76.0	74.8	77.9	85.6	78.6
✓	-	-	86.9	81.6	79.2	87.3	83.8
✓	✓	-	88.5	82.7	84.4	91.2	86.7
-	-	✓	77.4	79.4	82.6	89.9	82.3
✓	✓	✓	**89.1**	**84.5**	**87.8**	**93.8**	**88.8**

6 Related Work

6.1 Forensic Representations

Historically, forgery detection began with handcrafted representations, targeting blending boundary anomalies Li et al. (2020a); Nirkin et al. (2022); Shiohara & Yamasaki (2022) or up-sampling artifacts Tan et al. (2024b); Ma et al. (2025). However, their effectiveness is bound to specific priors, failing when these priors are violated. To overcome this limitation, researchers shifted towards data-driven deep representations, with Convolutional Neural Networks such as XceptionNet Chollet (2017); Cheng et al. (2025); Liu et al. (2025); Yan et al. (2023a); Qian et al. (2020); Zhang et al. (2025) and EfficientNet Tan & Le (2019); Yan et al. (2024a) becoming the backbones of detection frameworks. Although these models excel at extracting high-frequency anomalies Qian et al. (2020); Tan et al. (2024c) and model fingerprints Guillaro et al. (2023); Asnani et al. (2023), their inherent inductive bias towards local textures leads to overfitting specific local artifacts, resulting in limited generalization. Recently, a paradigm shift has emerged towards leveraging the universal capabilities of Vision-Language Models for forgery detection, predominantly adopting CLIP Radford et al. (2021); Shao et al. (2024); Guo et al. (2025) as the foundational architecture. Ojha et al. Ojha et al. (2023) pioneered this direction by revealing that the frozen CLIP representation inherently encapsulates generalizable artifacts, enabling the detection of fully synthetic images via simple linear probing. However, the efficacy of such frozen features remains limited when applied to specific facial manipulation tasks. To address this, subsequent research has explored Parameter-Efficient Fine-Tuning (PEFT) strategies to adapt CLIP Cui et al. (2025); Yan et al. (2025); Shao et al. (2024); Guo et al. (2025). For instance, Cui et al. Cui et al. (2025) introduced lightweight adapters and auxiliary convolutional branches. Zhou et al. (2026) exploit DINOv3's stronger alignment with diffusion traces to learn transferable forensic concepts. Unlike prior work that evaluates representations mainly through downstream accuracy, we use optimization stability as a representation-level diagnostic of whether frozen VFM features retain transferable forensic evidence.

6.2 Sharpness-Aware Minimization

Sharpness-Aware Minimization (SAM) (Foret et al., 2021) improves generalization by minimizing the worst-case loss within a Euclidean neighborhood. Its variants refine neighborhood geometry through Fisher information (Kim et al., 2022) or exponential tilting (Li et al., 2025), and improve optimization through Lookahead (Tan et al., 2024a) or sparse perturbations (Mi et al., 2025). Unified SAM (USAM) (Oikonomou & Loizou, 2025) further unifies update rules under a relaxed expected-residual condition, but does not explicitly characterize the data- and representation-dependent local stability threshold studied here. Whereas prior SAM studies primarily treat the perturbation radius as an optimizer design choice, we use the largest stable radius as a representation-level diagnostic and connect its analytical approximation to the GSNR.

7 Conclusion

This work identifies Optimization Collapse as a diagnostic manifestation of representational weakness in VFM-based forensics and introduces COR to characterize this vulnerability. Layer-wise analyses and controlled fidelity interpolation show that, as forgery fidelity increases, weak forensic cues are progressively attenuated in deeper layers. This attenuation is associated with lower GSNR and COR, less stable optimization, and poorer detector generalization. The COR-GSNR relationship shows that the squared COR is proportional to GSNR via a state-dependent statistical-geometric scaling factor; both theoretical and empirical analyses support the use of COR as an indicator of the generalizability of frozen VFM representations. Guided by this analysis, CoRIT combines regional refinement, progressive signal injection, and hierarchical representation integration to reduce gradient-proxy variance and preserve weak regional forensic cues across layers. Experiments demonstrate that CoRIT increases COR and GSNR, mitigates Optimization Collapse, and improves cross-manipulation and cross-domain generalization over VFM-based baselines while keeping the backbone frozen. Overall, our findings suggest that frozen VFM representations may not reliably preserve generalizable evidence for high-fidelity forgery detection. COR provides a useful diagnostic signal for this limitation, while CoRIT offers a practical way to mitigate the collapse without backbone adaptation.

REFERENCES

Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. Reverse engineering of generative models: Inferring model hyperparameters from generated images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):15477–15493, 2023.

Peter L. Bartlett, Philip M. Long, and Olivier Bousquet. The dynamics of sharpness-aware minimization: Bouncing across ravines and drifting towards wide minima. *J. Mach. Learn. Res.*, 24: 316:1–316:36, 2023.

benpflaum, Brian G, djdj, Irina Kofman, JE Tester, JLElliott, Joshua Metherd, Julia Elliott, Mozaic, Phil Culliton, Sohier Dane, and Woo Kim. Deepfake detection challenge. `https://kaggle.com/competitions/deepfake-detection-challenge`, 2019. Kaggle.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *CVPR*, pp. 4103–4112. IEEE, 2022.

Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? In *NeurIPS*, 2024.

Jikang Cheng, Ying Zhang, Qin Zou, Zhiyuan Yan, Chao Liang, Zhongyuan Wang, and Chen Li. Ed^4: Explicit data-level debiasing for deepfake detection. *IEEE Trans. Image Process.*, 34:4618–4630, 2025.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pp. 1800–1807. IEEE Computer Society, 2017.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting CLIP for generalizable face forgery detection. In *CVPR*, pp. 19207–19217. Computer Vision Foundation / IEEE, 2025.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*. OpenReview.net, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton-Ferrer. The deepfake detection challenge (DFDC) preview dataset. *CoRR*, abs/1910.08854, 2019.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*. OpenReview.net, 2021.

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

Xinghe Fu, Zhiyuan Yan, Taiping Yao, Shen Chen, and Xi Li. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *AAAI*, pp. 3040–3048. AAAI Press, 2025.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pp. 2232–2241. PMLR, 2019.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pp. 2672–2680, 2014.

Zhihao Gu, Taiping Yao, Yang Chen, Shouhong Ding, and Lizhuang Ma. Hierarchical contrastive inconsistency learning for deepfake video detection. In *ECCV (12)*, volume 13672 of *Lecture Notes in Computer Science*, pp. 596–613. Springer, 2022.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *CVPR*, pp. 20606–20615. IEEE, 2023.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *CVPR*, pp. 105–116. Computer Vision Foundation / IEEE, 2025.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *CVPR*, pp. 5039–5049. Computer Vision Foundation / IEEE, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *CVPR*, pp. 4490–4499. IEEE, 2023.

Hossein Kashiani, Niloufar Alipour Talemi, and Fatemeh Afghah. Freqdebias: Towards generalizable deepfake detection via consistency-driven frequency debiasing. In *CVPR*, pp. 8775–8785. Computer Vision Foundation / IEEE, 2025.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *ICLR*. OpenReview.net, 2017.

Minyoung Kim, Da Li, Shell Xu Hu, and Timothy M. Hospedales. Fisher SAM: information geometry and sharpness aware minimisation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11148–11161. PMLR, 2022.

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 *CVPR*, pp. 5000–5009. Computer Vision Foundation / IEEE, 2020a.

Tian Li, Tianyi Zhou, and Jeff A. Bilmes. Tilted sharpness-aware minimization. In *ICML*. OpenReview.net, 2025.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *CVPR*, pp. 3204–3213. Computer Vision Foundation / IEEE, 2020b.

Yuzhen Lin, Wentang Song, Bin Li, Yuezun Li, Jiangqun Ni, Han Chen, and Qiushi Li. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In *ECCV (86)*, volume 15144 of *Lecture Notes in Computer Science*, pp. 104–122. Springer, 2024.

Jinlong Liu, Yunzhi Bai, Guoqing Jiang, Ting Chen, and Huayan Wang. Understanding why neural networks generalize well through GSNR of parameters. In *ICLR*. OpenReview.net, 2020.

Jipeng Liu, Haichao Shi, Yaru Zhang, and Xiao-Yu Zhang. Knowledge negative distillation: Circumventing overfitting to unlock more generalizable deepfake detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 11318–11327, 2025.

Philip M. Long and Peter L. Bartlett. Sharpness-aware minimization and the edge of stability. *J. Mach. Learn. Res.*, 25:179:1–179:20, 2024.

Anwei Luo, Chenqi Kong, Jiwu Huang, Yongjian Hu, Xiangui Kang, and Alex C. Kot. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Trans. Inf. Forensics Secur.*, 19:1168–1182, 2024.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *CVPR*, pp. 16317–16326. Computer Vision Foundation / IEEE, 2021.

Long Ma, Zhiyuan Yan, Yize Chen, Jin Xu, Qinglang Guo, Hu Huang, Yong Liao, and Hui Lin. From specificity to generality: Revisiting generalizable artifacts in detecting face deepfakes. *CoRR*, abs/2504.04827, 2025.

James Martens. New insights and perspectives on the natural gradient method. *J. Mach. Learn. Res.*, 21:146:1–146:76, 2020.

Peng Mi, Li Shen, Tianhe Ren, Yiyi Zhou, Tianshuo Xu, Xiaoshuai Sun, Tongliang Liu, Rongrong Ji, and Dacheng Tao. Systematic investigation of sparse perturbed sharpness-aware minimization optimizer. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(10):8538–8549, 2025.

Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *CVPR*, pp. 17395–17405. IEEE, 2024.

Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. CORE: consistent representation learning for face forgery detection. In *CVPR Workshops*, pp. 12–21. IEEE, 2022.

Andrew Gully Nick Dufour. Deepfakedetection. `https://research.google/blog/contributing-data-to-deepfake-detection-research/`, 2019. Accessed: 2019 - 09 - 24.

Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6111–6121, 2022.

Dimitris Oikonomou and Nicolas Loizou. Sharpness-aware minimization: General analysis and improved rates. In *ICLR*. OpenReview.net, 2025.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, pp. 24480–24489. IEEE, 2023.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV (12)*, volume 12357 of *Lecture Notes in Computer Science*, pp. 86–103. Springer, 2020.

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 *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, pp. 1–11. IEEE, 2019.

Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

Rui Shao, Tianxing Wu, Jianlong Wu, Liqiang Nie, and Ziwei Liu. Detecting and grounding multimodal media manipulation and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5556–5574, 2024.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *CVPR*, pp. 18699–18708. IEEE, 2022.

Ke Sun, Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. In *NeurIPS*, 2024.

Chengli Tan, Jiangshe Zhang, Junmin Liu, and Yihong Gong. Sharpness-aware lookahead for accelerating convergence and improving generalization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):10375–10388, 2024a.

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 *CVPR*, pp. 28130–28139. IEEE, 2024b.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *AAAI*, pp. 5052–5060. AAAI Press, 2024c.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 2019.

Chao Xu, Jiangning Zhang, Yue Han, Guanzhong Tian, Xianfang Zeng, Ying Tai, Yabiao Wang, Chengjie Wang, and Yong Liu. Designing one unified framework for high-fidelity face reenactment and swapping. In *ECCV (15)*, volume 13675 of *Lecture Notes in Computer Science*, pp. 54–71. Springer, 2022.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. UCF: uncovering common features for generalizable deepfake detection. In *ICCV*, pp. 22355–22366. IEEE, 2023a.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *NeurIPS*, 2023b.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *CVPR*, pp. 8984–8994. IEEE, 2024a.

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 *NeurIPS*, 2024b.

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 *ICML*. OpenReview.net, 2025.

Zhongyi Zhang, Jie Zhang, Wenbo Zhou, Xinghui Zhou, Qing Guo, Weiming Zhang, Tianwei Zhang, and Nenghai Yu. Facetracer: Unveiling source identities from swapped face images and videos for fraud prevention. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(12):12021–12037, 2025.

Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *ICCV*, pp. 15003–15013. IEEE, 2021.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *ICCV*, pp. 15024–15034. IEEE, 2021.

Menyanshu Zhou, Ziyin Zhou, Ke Sun, Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. ForensicConcept: Transferable forensic concepts for AIGI detection. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Xiangyu Zhu, Hongyan Fei, Bin Zhang, Tianshuo Zhang, Xiaoyu Zhang, Stan Z. Li, and Zhen Lei. Face forgery detection by 3d decomposition and composition search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8342–8357, 2023.

A Overview of the Appendix

This appendix is organized into the following sections. The links below provide a concise guide to the contents.

- **Section B: Theoretical Analysis of Optimization Collapse.** This section defines Optimization Collapse and the Critical Optimization Radius (COR), derives a geometric stability bound for COR, and connects COR to the Gradient Signal-to-Noise Ratio (GSNR).
- **Section C: Derivation of the Contrastive Gradient Proxy.** This section derives the forward-computable feature residual used as the Contrastive Gradient Proxy (CGP), avoiding per-sample backpropagation through the frozen backbone.
- **Section D: Additional Experimental Details and Results.** This section presents datasets and evaluation protocols, additional generalization and efficiency results, theoretical visualizations, ablation studies, sensitivity analyses, and robustness evaluations.
- **Section E: Additional Evidence for the Fidelity–Generalization Link.** This section reports FID and text-semantic displacement analyses and introduces a controlled fidelity intervention to examine how forgery fidelity affects optimization and generalization.
- **Section F: Theoretical Derivations and Supporting Analysis.** This section provides the full Hessian–Covariance decomposition and a detailed justification for approximating local smoothness using the maximum Hessian eigenvalue.
- **Section G: Visualization of Region Refinement Masks Across Layers.** This section visualizes the Region Refinement Mask (RRM) throughout the backbone and analyzes the layer-wise distribution and suppression of outlier tokens.

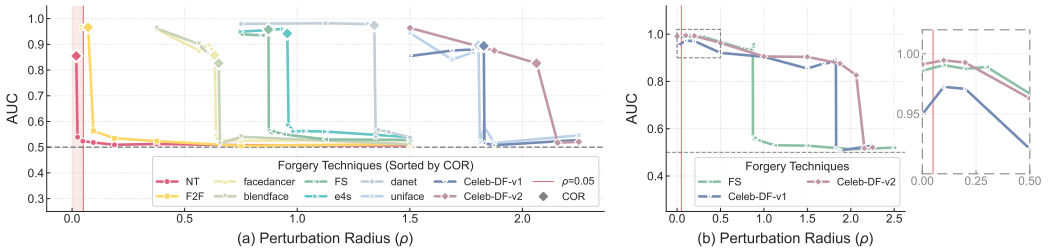


Figure 6: AUC curves of models trained with Sharpness-Aware Minimization (SAM) across varying perturbation radii ρ. (a) Empirical analysis of Optimization Collapse. The Critical Optimization Radius (COR) (diamond markers) across various datasets. The red region highlights optimization failure at the standard radius ($\rho = 0.05$). (b) AUC trends across the entire span of ρ. The gray dashed box indicates the localized region magnified on the right, highlighting AUC variations around the recommended radius ($\rho = 0.05$).

B Theoretical Analysis of Optimization Collapse: From Geometric Stability to Generalization

The empirical bifurcation observed in Figure 6 exposes an optimization failure phenomenon termed Optimization Collapse. To quantify this phenomenon, we first propose the Critical Optimization Radius (COR) and derive an analytical approximation from a sufficient geometric stability condition. Subsequently, we establish a theoretical link between this approximation and the Gradient Signal-to-Noise Ratio (GSNR) Liu et al. (2020), a metric of gradient signal quality associated with generalization. This connection isolates a statistical pathway through which degraded GSNR can contract the stability radius, without treating GSNR as its sole determinant. This theoretical insight guides the design of our proposed framework.

B.1 The Phenomenon of Optimization Collapse

Optimization Collapse occurs when the classifier's predictive capability drops to random guessing once the SAM perturbation radius ρ exceeds a narrow threshold. To investigate this, we systemati-

cally profiled AUC trajectories using linear probing on the frozen CLIP-ViT-L/14 backbone across ten representative face forgery paradigms. As Figure 6(a) illustrates, a consistent pattern emerges: when ρ exceeds its respective threshold, the AUC falls to approximately 0.5. We define this boundary as the Critical Optimization Radius (COR). A sufficiently large COR permits a wider range of stable SAM perturbations; however, a vanishing COR, as observed in NT and F2F, can place the SAM update outside the guaranteed stability regime.

Can simply reducing ρ resolve this collapse? We emphasize that SAM is employed here not as an indispensable optimizer, but as a controlled diagnostic probe exposing the intrinsic representational limitations of the backbone. As illustrated in Figure 6(b), increasing ρ from 0 to 0.1 yields consistent generalization gains, confirming the well-established benefits of landscape flattening. When $\rho = 0$, SAM degenerates to standard SGD, which converges stably but completely forfeits these generalization advantages. Naively reducing ρ circumvents the geometric instability yet simultaneously abandons the generalization potential, leaving the underlying representational deficit entirely unresolved. Our subsequent theoretical analysis formalizes this tension by linking the stability boundary (COR) to gradient signal quality (GSNR), ultimately reframing Optimization Collapse from a mere optimization failure into a profound representational diagnosis of intrinsic generalization capability.

B.2 Preliminaries and Definitions

We adopt standard notation to characterize the optimization dynamics. Let $\mathbf{w}_t$ denote the model parameters at step t. For the global objective function $\mathcal{L}(\mathbf{w})$, its value, gradient, and Hessian matrix at step t are denoted as $\mathcal{L}_t = \mathcal{L}(\mathbf{w}_t)$, $\nabla\mathcal{L}_t = \nabla\mathcal{L}(\mathbf{w}_t)$, and $\mathbf{H}_t = \nabla^2\mathcal{L}(\mathbf{w}_t)$, respectively. Let $\mathcal{B}_t$ denote the mini-batch sampled at step t. The stochastic gradient is defined as $\mathbf{g}_t = \nabla\mathcal{L}_{\mathcal{B}_t}(\mathbf{w}_t)$, which serves as an unbiased estimator of the global gradient, i.e., $\mathbb{E}_{\mathcal{B}_t}[\mathbf{g}_t] = \nabla\mathcal{L}_t$. Let $\tilde{\mathbf{g}}_t$ represent the SAM-perturbed gradient. To analyze the stability limits, we define ρ as the perturbation radius, ρ_{critical} as the Critical Optimization Radius (COR), and t^* as the Bottleneck Step. We quantify the local geometry via the local smoothness constant L_t. $\mathbb{E}[\cdot]$ and $\mathrm{Cov}[\cdot]$ denote the expectation and covariance operators; $\mathrm{Tr}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the trace and the maximum eigenvalue of a square matrix, respectively; and $\|\cdot\|$ represents the Euclidean norm.

For ease of reference, Table 4 summarises the principal symbols used throughout this appendix.

B.2.1 Sharpness-Aware Minimization (SAM)

Standard SAM seeks parameters lying within a flat loss landscape to enhance generalization by solving the following minimax objective Foret et al. (2021):

$$\min_{\mathbf{w}} \max_{\|\boldsymbol{\epsilon}\|_2 \le \rho} \mathcal{L}(\mathbf{w} + \boldsymbol{\epsilon}), \tag{12}$$

where $\rho \ge 0$ represents a pre-defined perturbation radius. Specifically, at step t, SAM computes the perturbed gradient $\tilde{\mathbf{g}}_t = \nabla\mathcal{L}(\mathbf{w}_t + \boldsymbol{\epsilon}_t)$ via an optimal adversarial perturbation vector $\boldsymbol{\epsilon}_t$. As derived by Foret et al. Foret et al. (2021) via a first-order Taylor approximation, this optimal adversarial perturbation aligns with the normalized gradient direction:

$$\boldsymbol{\epsilon}_t \approx \rho\mathbf{g}_t/\|\mathbf{g}_t\|_2. \tag{13}$$

By penalizing sharp minima, as illustrated in Figure 6(b), SAM yields consistent generalization gains at a recommended radius ($\rho \approx 0.05$). As ρ further increases, the generalization performance gradually degrades, until exceeding the COR triggers Optimization Collapse.

B.2.2 Geometric Stability Condition

We observe that SAM obtains the perturbed gradient $\tilde{\mathbf{g}}_t$ by injecting an adversarial perturbation $\boldsymbol{\epsilon}_t$ upon the stochastic gradient $\mathbf{g}_t$. For effective optimization, the intensity of this adversarial perturbation must not overwhelm the original gradient signal. We quantify this impact via the expected inner product between the gradient $\nabla\mathcal{L}_t$ and the perturbed update $\tilde{\mathbf{g}}_t$. The inner product captures both directional similarity and magnitude similarity, where an excessive perturbation radius ρ renders the descent ineffective by disrupting this similarity. To guarantee stable descent, we formulate a geometric stability constraint: the SAM update direction $-\tilde{\mathbf{g}}_t$ must, in expectation, maintain a positive projection onto the true descent direction $-\nabla\mathcal{L}_t$, equivalently:

$$\mathbb{E}[\langle\nabla\mathcal{L}_t, \tilde{\mathbf{g}}_t\rangle] > 0. \tag{14}$$

Table 4: Summary of principal notation.

Symbol	Definition
Optimization & Loss	
$\mathbf{w}_t$	Model parameters at optimization step t
$\mathcal{L}_t$	Expected NLL loss $\mathbb{E}_{\mathbf{x}\sim q}[-\log p_{\mathbf{w}_t}(\mathbf{x})]$
$\mathbf{g}(\mathbf{x}, \mathbf{w}_t)$	Stochastic gradient for a single sample $\mathbf{x}$
$\nabla\mathcal{L}_t$	Expected (population) gradient
$\mathbf{H}_t$	Expected Hessian matrix $\nabla^2\mathcal{L}(\mathbf{w}_t)$
L_t	Local smoothness constant within the ρ-neighborhood
t^*	Bottleneck optimization step
Hessian Decomposition	
$\mathrm{Cov}[\mathbf{g}_t]$	Gradient covariance matrix
$\mathbf{\Xi}_t$	Model misspecification residual $\mathbb{E}_q\big[\nabla^2 p_{\mathbf{w}_t}/p_{\mathbf{w}_t}\big]$
γ_t	State-dependent statistical-geometric scaling factor $\mathrm{Tr}(\mathrm{Cov}[\mathbf{g}_t])/\lambda^2_{\max}(\mathbf{H}_t)$
$\lambda_{\max}(\mathbf{H}_t)$	Largest eigenvalue of $\mathbf{H}_t$
Core Indicators	
GSNR_{t^*}	Gradient Signal-to-Noise Ratio at t^*: $\|\nabla\mathcal{L}_{t^*}\|^2/\,\mathrm{Tr}(\mathrm{Cov}[\mathbf{g}_{t^*}])$
ρ_{critical}	Critical Optimization Radius
Distributions	
$q(\mathbf{x})$	True data distribution
$p_{\mathbf{w}_t}(\mathbf{x})$	Parameterised model distribution at step t

This inequality serves as the theoretical foundation for deriving the COR.

B.2.3 Gradient Signal-to-Noise Ratio (GSNR)

Introduced by Liu et al. Liu et al. (2020) to quantify generalization capability, GSNR measures the strength of the true gradient signal relative to the stochastic gradient variance. To align with the scalar perturbation radius ρ, we define the global GSNR at step t as:

$$\mathrm{GSNR}_t = \frac{\|\nabla\mathcal{L}_t\|^2}{\mathrm{Tr}(\mathrm{Cov}[\mathbf{g}_t])}. \tag{15}$$

A lower GSNR indicates a degraded gradient signal, which is theoretically linked to poor generalization.

B.2.4 Hessian-Covariance Decomposition

Second-order optimization theory Martens (2020) establishes a correspondence between the Hessian's geometric curvature and the gradient covariance's statistical properties. Under the negative log-likelihood (NLL) loss, let $\mathbf{x} \sim q(\mathbf{x})$ denote the input data and $p_{\mathbf{w}_t}(\mathbf{x})$ be the parameterized model distribution. The stochastic gradient is formulated as $\mathbf{g}_t = -\nabla_{\mathbf{w}_t}\log p_{\mathbf{w}_t}(\mathbf{x})$. The expected Hessian matrix $\mathbf{H}_t$ decomposes as follows:

$$\mathbf{H}_t = \mathrm{Cov}[\mathbf{g}_t] + \nabla\mathcal{L}_t\nabla\mathcal{L}_t^\top - \mathbf{\Xi}_t, \tag{16}$$

where $\mathrm{Cov}[\mathbf{g}_t]$ is the gradient covariance matrix, and $\mathbf{\Xi}_t = \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}\left[\frac{\nabla^2_{\mathbf{w}_t} p_{\mathbf{w}_t}(\mathbf{x})}{p_{\mathbf{w}_t}(\mathbf{x})}\right]$ is the residual term for model misspecification. The details are provided in the appendix.

B.3 Assumptions

B.3.1 Optimization Setting

We focus our theoretical analysis on the optimization process trained via mini-batch SGD with Cross-Entropy (CE) loss. For classification task, CE loss coincides with the Negative Log-Likelihood (NLL), which validates the Hessian-Covariance Decomposition in Section B.2.4.

Assumption 1 (Local Curvature & Mini-batch Smoothness)**.** *Assume the loss function $\mathcal{L}(\mathbf{w})$ is twice continuously differentiable within the ρ-neighborhood $\mathcal{N}(\mathbf{w}_t, \rho) = \{\mathbf{w} \mid \|\mathbf{w} - \mathbf{w}_t\|_2 \leq \rho\}$. We define the local smoothness constant L_t via the supremum of the Hessian's spectral norm, which approximates its maximum eigenvalue for small ρ:*

$$L_t = \sup_{\mathbf{u} \in \mathcal{N}(\mathbf{w}_t, \rho)} \|\nabla^2 \mathcal{L}(\mathbf{u})\|_2 \approx \lambda_{\max}(\mathbf{H}_t). \tag{17}$$

We assume this L_t-Lipschitz smoothness holds uniformly across mini-batches. For any perturbation $\|\boldsymbol{\epsilon}\| \leq \rho$:

$$\|\tilde{\mathbf{g}}_t - \mathbf{g}_t\| = \|\nabla \mathcal{L}_{\mathcal{B}_t}(\mathbf{w}_t + \boldsymbol{\epsilon}) - \nabla \mathcal{L}_{\mathcal{B}_t}(\mathbf{w}_t)\| \leq L_t \|\boldsymbol{\epsilon}\|. \tag{18}$$

Justification. *The approximation in Eq.* (17) *and the mini-batch smoothness assumption are standard in SAM theoretical analysis Foret et al. (2021); Bartlett et al. (2023); Long & Bartlett (2024). Although the bound may become loose for large perturbation radii, it remains tight in the* Optimization Collapse *regime where the critical radius is vanishingly small (*$\mathrm{COR} < 0.05$*). In such narrow neighborhoods, the maximum eigenvalue of the Hessian exhibits negligible variation, ensuring the approximation's validity.*

B.4 Derivation of the Critical Optimization Radius

To obtain an analytical approximation to the empirical COR, we translate the geometric stability condition (Section B.2.2) into a quantifiable boundary. This yields a step-wise upper bound for the perturbation radius ρ that guarantees descent.

Lemma 1 (Local Stability Bound)**.** *Under local L_t-Lipschitz smoothness (Assumption 1), the geometric stability condition $\mathbb{E}[\langle \nabla \mathcal{L}_t, \tilde{\mathbf{g}}_t \rangle] > 0$ holds if:*

$$\rho < \frac{\|\nabla \mathcal{L}_t\|}{L_t}. \tag{19}$$

Proof. Decomposing the expected inner product in Eq. (14) yields:

$$\begin{aligned}
\mathbb{E}[\langle \nabla \mathcal{L}_t, \tilde{\mathbf{g}}_t \rangle] &= \mathbb{E}[\langle \nabla \mathcal{L}_t, \mathbf{g}_t \rangle] + \mathbb{E}[\langle \nabla \mathcal{L}_t, \tilde{\mathbf{g}}_t - \mathbf{g}_t \rangle] \\
&= \|\nabla \mathcal{L}_t\|^2 + \mathbb{E}[\langle \nabla \mathcal{L}_t, \tilde{\mathbf{g}}_t - \mathbf{g}_t \rangle] && \text{(Unbiasedness)} \\
&\geq \|\nabla \mathcal{L}_t\|^2 - \mathbb{E}[\|\nabla \mathcal{L}_t\| \|\tilde{\mathbf{g}}_t - \mathbf{g}_t\|] && \text{(Cauchy-Schwarz)} \\
&\overset{(18)}{\geq} \|\nabla \mathcal{L}_t\|^2 - \mathbb{E}[\|\nabla \mathcal{L}_t\| (L_t \|\boldsymbol{\epsilon}_t\|)] \\
&\overset{(13)}{\geq} \|\nabla \mathcal{L}_t\|^2 - L_t \rho \|\nabla \mathcal{L}_t\|.
\end{aligned} \tag{20}$$

For Eq. (14) to hold, we require $\|\nabla \mathcal{L}_t\|^2 - L_t \rho \|\nabla \mathcal{L}_t\| > 0$. Rearranging this inequality yields:

$$\rho < \frac{\|\nabla \mathcal{L}_t\|}{L_t}. \tag{21}$$

□

Lemma 1 provides a step-wise stability bound. However, the Critical Optimization Radius (COR) from Section B.1 acts as a global threshold throughout the optimization process, remaining independent of step t. The COR is dictated by the most restrictive local stability bound, attained at what we identify as the bottleneck step t^*.

Definition 1 (Critical Optimization Radius)**.** *We define the analytical COR approximation $\rho_{critical}$ as the minimum local stability bound over the optimization trajectory. Applying the curvature approximation $L_t \approx \lambda_{\max}(\mathbf{H}_t)$ from Assumption 1, it is formulated as:*

$$\rho_{critical} = \min_t \frac{\|\nabla \mathcal{L}_t\|}{\lambda_{\max}(\mathbf{H}_t)}. \tag{22}$$

We define the bottleneck step t^* as the iteration achieving this minimum, where the local curvature restricts the perturbation radius the most:

$$t^* = \operatorname*{argmin}_t \frac{\|\nabla \mathcal{L}_t\|}{\lambda_{\max}(\mathbf{H}_t)}. \tag{23}$$

Because Lemma 1 provides a sufficient rather than necessary descent condition, ρ_{critical} is used as an analytical approximation to the empirical collapse threshold measured by the ρ-sweep, not as an exact boundary. Evaluating it at the bottleneck step t^* yields $\rho_{\text{critical}} = \|\nabla \mathcal{L}_{t^*}\|/\lambda_{\max}(\mathbf{H}_{t^*})$.

B.5 Bridging COR and Generalization via GSNR

While COR is derived from geometric constraints, we establish a formal connection between this stability threshold and the Gradient Signal-to-Noise Ratio (GSNR). By decomposing the squared COR at the bottleneck step, we demonstrate that SAM's geometric stability couples directly with the statistical quality of the gradient.

Theorem 1 (COR–GSNR Relationship)**.** *The squared analytical COR approximation $\rho^2_{critical}$ at the bottleneck step t^* decomposes as:*

$$\rho^2_{critical} = \gamma_{t^*} \cdot GSNR_{t^*}, \qquad \gamma_{t^*} = \frac{\mathrm{Tr}(\mathrm{Cov}[\mathbf{g}_{t^*}])}{\lambda^2_{\max}(\mathbf{H}_{t^*})}, \tag{24}$$

where γ_{t^*} is the state-dependent statistical-geometric scaling factor linking stochastic gradient fluctuations to local sharpness.

Proof. By the definitions of COR and GSNR:

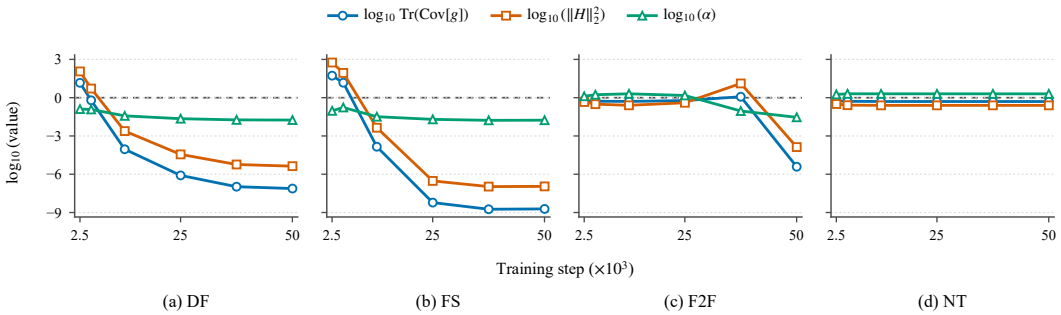
$$\begin{aligned} \rho^2_{\text{critical}} &\overset{(22)}{=} \frac{\|\nabla \mathcal{L}_{t^*}\|^2}{\lambda^2_{\max}(\mathbf{H}_{t^*})} \\ &\overset{(15)}{=} \frac{\mathrm{Tr}(\mathrm{Cov}[\mathbf{g}_{t^*}]) \cdot \mathrm{GSNR}_{t^*}}{\lambda^2_{\max}(\mathbf{H}_{t^*})} \\ &= \frac{\mathrm{Tr}(\mathrm{Cov}[\mathbf{g}_{t^*}])}{\lambda^2_{\max}(\mathbf{H}_{t^*})} \cdot \mathrm{GSNR}_{t^*} = \gamma_{t^*} \cdot \mathrm{GSNR}_{t^*}. \end{aligned} \tag{25}$$

□

Remark 1. The coefficient γ_{t^*} is a state-dependent statistical-geometric scaling factor. The numerator $\mathrm{Tr}(\mathrm{Cov}[\mathbf{g}_{t^*}])$ is the total gradient noise energy across all parameter directions, while the denominator $\lambda^2_{\max}(\mathbf{H}_{t^*})$ is the squared curvature along the sharpest direction. They jointly reflect the local parameter state. The numerator characterizes local instability from a gradient-statistical perspective, whereas the denominator characterizes it from a curvature-geometric perspective. Thus, γ_{t^*} serves as a statistical-geometric scaling factor between stochastic gradient fluctuations and local sharpness. When γ_{t^*} remains approximately stable across settings, the COR ranking is governed by GSNR alone.

B.5.1 Empirical Observation of γ_t Stability

Figure 7 shows the trajectories of the numerator $\mathrm{Tr}(\mathrm{Cov}[\mathbf{g}_t])$, the denominator $\|\mathbf{H}_t\|_2^2$, and γ_t during training step t. The numerator and denominator of γ_t change together, keeping γ_t substantially more stable than either raw scale term in the trajectories relevant to the analysis. The conclusion is that γ_t is a state-dependent statistical-geometric scaling factor with limited variation, so the cross-dataset and layer-wise variation of the COR is primarily driven by GSNR.



(a) DF (b) FS (c) F2F (d) NT

Figure 7: Trajectories of the gradient covariance trace, squared Hessian spectral norm, and γ_t during training step t. Values are shown on a $\log_{10}$ scale for DF, FS, F2F, and NT.

B.5.2 A Heuristic Explanation for the Approximate Stability of γ_t

The Hessian–covariance decomposition in Eq. (16) provides a heuristic explanation for why γ_t may remain approximately stable across different settings. Taking the trace of both sides gives

$$\mathrm{Tr}(\mathbf{H}_t) = \mathrm{Tr}(\mathrm{Cov}[\mathbf{g}_t]) + \|\nabla \mathcal{L}_t\|^2 - \mathrm{Tr}(\mathbf{\Xi}_t). \tag{26}$$

Let

$$V_t = \mathrm{Tr}(\mathrm{Cov}[\mathbf{g}_t]) \quad \text{and} \quad S_t = \|\nabla \mathcal{L}_t\|^2.$$

Rearranging Eq. (26) yields the exact relation

$$V_t = \mathrm{Tr}(\mathbf{H}_t) + \mathrm{Tr}(\mathbf{\Xi}_t) - S_t. \tag{27}$$

At the bottleneck step t^*, the GSNR is empirically small, so that $S_{t^*}/V_{t^*} \ll 1$. Consequently, the contribution of the signal term in Eq. (27) is relatively small, and we obtain the approximation

$$V_{t^*} \approx \mathrm{Tr}(\mathbf{H}_{t^*}) + \mathrm{Tr}(\mathbf{\Xi}_{t^*}). \tag{28}$$

Substituting this approximation into the definition of γ_{t^*} gives

$$\gamma_{t^*} \approx \frac{\mathrm{Tr}(\mathbf{H}_{t^*}) + \mathrm{Tr}(\mathbf{\Xi}_{t^*})}{\lambda^2_{\max}(\mathbf{H}_{t^*})}. \tag{29}$$

This relation suggests that, in the low-GSNR regime, the dominant contribution to γ_{t^*} may come from the local curvature and the model–data mismatch captured by $\mathbf{\Xi}_{t^*}$, rather than directly from the magnitude of the gradient signal.

The proposed interpretation is instead that the curvature-related quantities in Eq. (29) may vary more slowly across datasets than the data-dependent gradient signal. Under this empirical condition, γ_{t^*} can appear approximately stable, whereas GSNR may vary substantially between high- and low-fidelity forgeries. This hypothesis can be tested by measuring the cross-dataset variation of $\mathrm{Tr}(\mathbf{H}_{t^*})$, $\mathrm{Tr}(\mathbf{\Xi}_{t^*})$, and $\lambda^2_{\max}(\mathbf{H}_{t^*})$ separately.

B.6 Theoretical Insights

Theorem 1 provides a direct interpretation of Optimization Collapse: since $\rho^2_{\text{critical}} = \gamma_{t^*} \cdot \mathrm{GSNR}_{t^*}$, when γ_{t^*} remains relatively stable, a vanishing COR (e.g., $\rho_{\text{critical}} \approx 0.001$ on NT) is associated with diminished GSNR, where the directional gradient signal is weak relative to gradient variance. Therefore, rather than only reducing the hyperparameter ρ of SAM optimizer, the relationship motivates improving gradient signal quality.

This theoretical insight also offers a guideline for our methodology design: by improving GSNR, we aim to lift the COR beyond the collapse regime ($\rho_{\text{critical}} < 0.05$), thereby mitigating Optimization Collapse. In practice, this equates to enhancing the strength and suppressing the variance of gradient signals.

C Derivation of the Contrastive Gradient Proxy

Maximizing the COR necessitates elevating the GSNR of the model parameters $\mathbf{w}$. However, directly maximizing $\mathrm{GSNR}(\mathbf{w})$ as an auxiliary optimization objective is computationally impractical: it requires per-sample backpropagation, which imposes an untenable overhead for large-scale architectures such as CLIP-ViT-Large. To circumvent this computational bottleneck, we shift our analytical focus from the parameter space to the feature space.

Let $\mathbf{V}^{(l)} \in \mathbb{R}^{N\times D}$ denote the visual tokens at layer l, let $h(\cdot)$ be the classifier head that maps these tokens to a predictive logit, and write the sample loss as $\mathcal{L} = \ell(h(\mathbf{V}^{(l)}), y)$. Denoting the feature Jacobian by $\mathbf{J}_{\mathbf{w}}^{(l)} = \partial \,\mathrm{vec}(\mathbf{V}^{(l)})/\partial \mathbf{w}$, the chain rule gives

$$\begin{aligned}\nabla_{\mathbf{w}}\mathcal{L} &= \left(\mathbf{J}_{\mathbf{w}}^{(l)}\right)^{\top} \nabla_{\mathrm{vec}(\mathbf{V}^{(l)})}\mathcal{L} \\ &= \frac{\partial \ell}{\partial h}\left(\mathbf{J}_{\mathbf{w}}^{(l)}\right)^{\top} \nabla_{\mathrm{vec}(\mathbf{V}^{(l)})} h.\end{aligned} \tag{30}$$

Eq. (30) makes explicit that the feature-gradient direction is the signal mapped by the network Jacobian into parameter space. It therefore motivates constructing a feature-space surrogate before this Jacobian mapping.

Inspired by the Self-Blended Image (SBI) paradigm (Shiohara & Yamasaki, 2022), we consider a paired original image and its self-blended counterpart. Their layer-l feature residual is

$$\Delta\mathbf{V}^{(l)} = \mathbf{V}_{\mathrm{SBI}}^{(l)} - \mathbf{V}_{\mathrm{orig}}^{(l)}. \tag{31}$$

Taking h to be the fake-class logit, SBI-based discrimination encourages a positive logit discrepancy

$$\Delta h = h\Big(\mathbf{V}_{\mathrm{SBI}}^{(l)}\Big) - h\Big(\mathbf{V}_{\mathrm{orig}}^{(l)}\Big). \tag{32}$$

A first-order Taylor expansion around $\mathbf{V}_{\mathrm{orig}}^{(l)}$ yields

$$\Delta h = \Big\langle \nabla_{\mathbf{V}^{(l)}} h\Big(\mathbf{V}_{\mathrm{orig}}^{(l)}\Big), \Delta\mathbf{V}^{(l)} \Big\rangle_F + \mathcal{O}\Big(\|\Delta\mathbf{V}^{(l)}\|_F^2\Big), \tag{33}$$

where $\langle\cdot,\cdot\rangle_F$ and $\|\cdot\|_F$ denote the Frobenius inner product and norm, respectively. For a fixed residual magnitude, the Cauchy–Schwarz inequality gives

$$\Big\langle \nabla_{\mathbf{V}^{(l)}} h, \Delta\mathbf{V}^{(l)} \Big\rangle_F \le \|\nabla_{\mathbf{V}^{(l)}} h\|_F \left\|\Delta\mathbf{V}^{(l)}\right\|_F, \tag{34}$$

with equality if and only if the two quantities are positively collinear. Thus, maximizing the first-order SBI logit discrepancy promotes

$$\frac{\nabla_{\mathbf{V}^{(l)}} h}{\|\nabla_{\mathbf{V}^{(l)}} h\|_F} \approx \frac{\Delta\mathbf{V}^{(l)}}{\|\Delta\mathbf{V}^{(l)}\|_F}. \tag{35}$$

Substituting this directional relation into Eq. (30) shows the explicit connection from the parameter gradient to the feature residual:

$$\nabla_{\mathbf{w}}\mathcal{L} \approx \beta^{(l)} \left(\mathbf{J}_{\mathbf{w}}^{(l)}\right)^{\top} \mathrm{vec}\Big(\Delta\mathbf{V}^{(l)}\Big), \tag{36}$$

where $\beta^{(l)}$ absorbs the loss derivative and the relative magnitude of the feature gradient. Eq. (36) describes the gradient that would be obtained by differentiating through the backbone. CoRIT instead freezes the backbone; explicitly forming this Jacobian or storing per-sample gradients would defeat the purpose of an efficient proxy. We therefore retain the forward-computable feature-space direction in Eq. (31) and define the Contrastive Gradient Proxy (CGP) as

$$\mathbf{g}_{\mathrm{CGP}}^{(l)} = \Delta\mathbf{V}^{(l)} = \mathbf{V}_{\mathrm{SBI}}^{(l)} - \mathbf{V}_{\mathrm{orig}}^{(l)}. \tag{37}$$

This proxy requires only two forward passes and eliminates both per-sample backpropagation and the memory overhead of gradient storage.

Table 5: Cross-domain performance on classical forensic benchmarks (Video-Level AUC (%)). Best results are in **bold**.

Method	Venues	CDFv2	DFDC	DFDCP	Avg.
LipForensics (Haliassos et al., 2021)	CVPR'21	82.4	73.5	-	-
FTCN (Zheng et al., 2021)	ICCV'21	86.9	-	74.0	-
PCL + I2G (Zhao et al., 2021)	ICCV'21	90.0	67.5	74.3	77.3
HCIL (Gu et al., 2022)	ECCV'22	79.0	69.2	-	-
SBI (Shiohara & Yamasaki, 2022)	CVPR'22	92.8	71.9	85.5	83.4
3DDet. (Zhu et al., 2023)	TPAMI'23	72.0	72.4	-	-
LSDA (Yan et al., 2024a)	CVPR'24	91.1	77.0	-	-
CFM (Luo et al., 2024)	TIFS'24	89.7	-	80.2	-
LAA-Net (Nguyen et al., 2024)	CVPR'24	84.0	-	74.1	-
FIA-USA (Ma et al., 2025)	NIPS'25	83.9	74.3	-	-
Effort (Yan et al., 2025)	ICML'25	**95.6**	84.3	90.9	90.3
Ours	-	94.1	**87.6**	**91.3**	**91.0**

D Additional Experimental Details and Results

D.1 Datasets and Evaluation Protocols

D.1.1 Classical Forensic Benchmarks.

We utilize the FaceForensics++ (FF++) (Rössler et al., 2019) dataset, which contains four manipulation types: Identity Replacement (DeepFakes, FaceSwap), Expression Reenactment (Face2Face), and Neural Rendering (NeuralTextures). For cross-manipulation evaluation, the model is trained on three manipulation types and tested on the held-out type. For cross-dataset evaluation, we train on FF++ and test on several widely used datasets: Deepfake Detection Challenge (DFDC) (benpflaum et al., 2019), its preview version (DFDCP) (Dolhansky et al., 2019), DeepFake Detection (DFD) (Nick Dufour, 2019), and both iterations of the Celeb-DF benchmark (CDF-v1 and CDF-v2) (Li et al., 2020b). All data preprocessing and evaluation protocols follow the standards of DeepfakeBench (Yan et al., 2023b) to ensure reproducibility and fair comparison. We report Frame-level and Video-level AUC (Area Under the Curve) as the primary metrics.

D.1.2 Next-Generation Generative Benchmarks.

To further test against rapidly evolving face forgery techniques, we incorporate the DF40 benchmark (Yan et al., 2024b), a large-scale benchmark covering 40 distinct generation methods. Following prior work (Ma et al., 2025), we select representative face forgery types and adopt Protocol-2 (Cross-Domain Evaluation) on the CDF domain. We report Frame-level AUC as the evaluation metric.

D.2 Generalization Evaluation

D.2.1 Generalization on Classical Forensic Benchmarks

In the video-level evaluation (Table 5), CoRIT achieves a state-of-the-art average AUC of 91.0%, outperforming the LoRA-based PEFT competitor Effort.

D.2.2 Generalization towards Emerging Forgeries

The generalization superiority of CoRIT extends to a broader spectrum of emerging facial forgery techniques and natural scene synthesis. On the Next-Generation Generative Benchmarks (Table 6), CoRIT consistently outperforms the CNN-based FIA-USA by 4.4% on average and improves over the frozen-backbone baseline UniFD by 8.8%, confirming its robustness against rapidly evolving generation techniques.

Table 6: Cross-domain performance on five representative face forgery types in DF40 (Frame-Level AUC (%)). Best results are in **bold**.

Method	Venues	uniface	e4s	facedancer	fsgan	inswap	Avg.
Recce (Cao et al., 2022)	CVPR'22	84.2	65.2	78.3	88.4	79.5	79.1
SBI (Shiohara & Yamasaki, 2022)	CVPR'22	64.4	69.0	44.7	87.9	63.3	65.9
CORE (Ni et al., 2022)	CVPRW'22	81.7	63.4	71.7	91.1	79.4	77.5
IID (Huang et al., 2023)	CVPR'23	79.5	71.0	79.0	86.4	74.4	78.1
UCF (Yan et al., 2023a)	ICCV'23	78.7	69.2	80.0	88.1	76.8	78.6
LSDA (Yan et al., 2024a)	CVPR'24	85.4	68.4	75.9	83.2	81.0	78.8
CDFA (Lin et al., 2024)	ECCV'24	76.5	67.4	75.4	84.8	72.0	75.2
ProgressiveDet (Cheng et al., 2024)	NIPS'24	84.5	71.0	73.6	86.5	78.8	78.9
FIA-USA (Ma et al., 2025)	NIPS'25	**91.8**	87.5	83.0	86.3	87.4	87.2
UniFD (Ojha et al., 2023)	CVPR'23	82.7	91.4	77.9	85.6	76.5	82.8
Ours	-	**91.8**	**94.4**	**89.2**	**94.2**	**88.4**	**91.6**

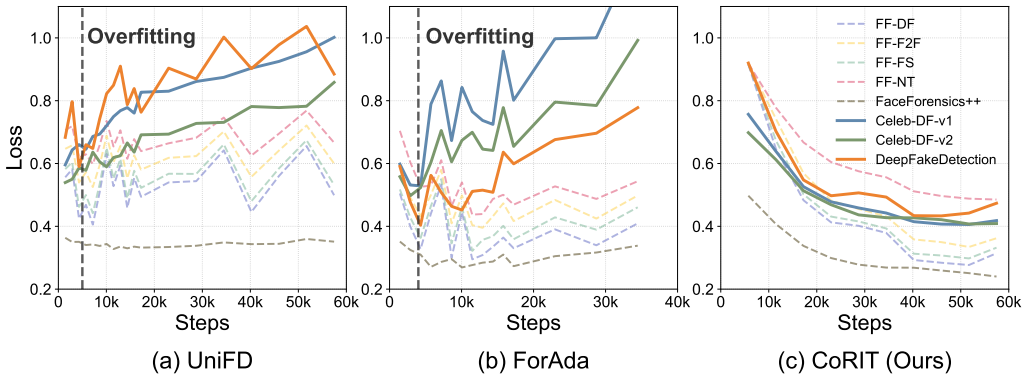


Figure 8: Training loss curves of compared methods illustrating generalization stability. Dashed lines indicate intra-dataset evaluation; solid lines indicate cross-dataset evaluation.

D.2.3 Generalization Stability

CoRIT shows strong generalization stability during training. As shown in Figure 8, competing methods exhibit a typical 'U-shaped' overfitting curve, whereas CoRIT maintains consistent performance under prolonged training. This indicates that our design encourages learning domain-invariant forensic representations, reducing the reliance on early stopping to avoid overfitting.

D.2.4 Computational Efficiency

CoRIT is highly memory-efficient during the training phase. Since the backbone remains entirely frozen, no gradient storage is needed for its parameters, reducing the overall GPU memory consumption from 21,224 MB (ForAda) to 2,130 MB, achieving a 90% reduction.

D.3 Theoretical Verification and Visualization

D.3.1 Loss Landscape Visualization

Figure 9 shows the loss landscapes of the final-layer (Layer 24) representations from UniFD and CoRIT, evaluated on a lightweight learnable classifier. CoRIT produces a flatter and smoother landscape, which is generally associated with better generalization (Keskar et al., 2017; Foret et al., 2021). The landscape for UniFD, particularly on the NT subset, is irregular with sharp curvature in multiple regions. As established in Lemma 1, such high curvature leads to reduced COR values, consistent with the empirical results in Figure 5 of the main paper.

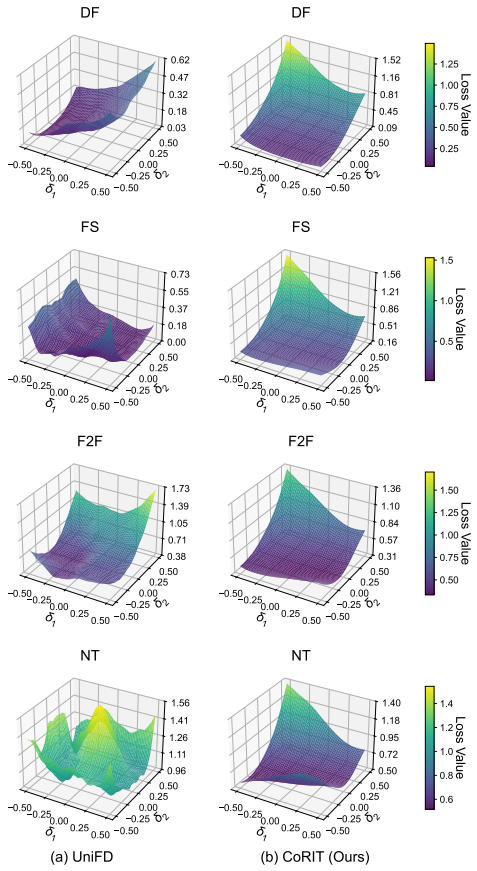


Figure 9: Comparison of loss landscapes between UniFD and the proposed CoRIT on the NT dataset.

D.4 Additional Ablation and Robustness Results

D.4.1 Ablation on Spatial Regions and Pooling

Table 7 presents the ablation on region configurations, including the number of regions K and pooling strategies. With a single region ($K = 1$), the foreground captures the most discriminative cues at 86.8% average AUC, which significantly outperforms random sampling. Two-region combinations ($K = 2$), such as foreground paired with boundary, raise performance to 88.4%. Using all three regions ($K = 3$) brings further gains on CDFv1 and DFDC, yielding a total improvement of 2.1% over foreground alone, which confirms the complementary value of boundary and background features. For pooling strategies, average pooling outperforms max pooling by 4.8% under the full-context setting, suggesting it better suppresses domain-specific noise that hinders generalization.

D.5 Sensitivity Analysis

D.5.1 Impact of Backbones

Table 8 evaluates CoRIT across backbone scales and pre-training paradigms, including ImageNet (Deng et al., 2009) and CLIP. CoRIT consistently outperforms UniFD under all backbones, with a 6.2% gain on CLIP-pretrained ViT-Base and 10.6% on ImageNet-pretrained ViT-L. These results show that our framework is robust to model scale and extends beyond multi-modal architectures to benefit backbones pre-trained for semantic classification. We adopt CLIP-ViT-L as the default backbone given its best overall performance.

Table 7: Ablation study on spatial regions and pooling strategies (Frame-Level AUC (%)). F, Bd, and Bg denote Foreground, Boundary, and Background, respectively. K is the number of regions. Best results are in **bold**.

Region	K	Pooling	CDFv1	CDFv2	DFDCP	DFDC	DFD	Avg.
Random	3	Avg.	76.1	79.7	80.9	78.6	88.0	80.7
F	1	Avg.	88.6	86.6	84.9	82.5	91.2	86.8
Bd	1	Avg.	76.4	77.3	78.1	78.2	87.4	79.5
Bg	1	Avg.	67.1	61.7	70.5	70.0	77.0	69.3
F+Bd	2	Avg.	89.3	**89.1**	88.5	83.7	**91.3**	88.4
F+Bg	2	Avg.	89.9	**89.1**	87.3	83.2	**91.3**	88.2
Bd+Bg	2	Avg.	77.8	81.2	79.8	78.0	88.0	81.0
F+Bd+Bg	3	Max	83.4	83.6	84.5	79.7	89.5	84.1
F+Bd+Bg	3	Avg.	**90.9**	**89.1**	**88.8**	**84.5**	**91.3**	**88.9**

Table 8: Ablation study across different backbones: IMG-L (ImageNet-ViT-Large), CLIP-B (CLIP-ViT-Base), and CLIP-L (CLIP-ViT-Large). Best results are in **bold**.

Backbone	Method	CDFv1	CDFv2	DFDCP	DFDC	Avg.
IMG-L	UniFD (Ojha et al., 2023)	53.0	54.0	63.2	61.4	57.9
	Ours	66.3	65.0	75.0	67.6	68.5
CLIP-B	UniFD (Ojha et al., 2023)	75.0	73.3	72.4	70.0	72.7
	Ours	81.0	77.2	80.8	76.5	78.9
CLIP-L	UniFD (Ojha et al., 2023)	71.2	76.0	73.6	74.8	73.9
	Ours	**90.9**	**89.1**	**88.8**	**84.5**	**88.3**

D.5.2 Impact of Mask Threshold

Table 9 shows that performance remains stable across the range $\tau \in [0, 0.9]$, demonstrating the model's robustness to this hyperparameter. We adopt $\tau = 0.1$ for all experiments.

D.5.3 Impact of Fusion Layers

Table 10 shows that performance remains stable across the range $l_{mid} \in [16, 20]$, indicating low sensitivity to the choice of fusion layer. We set $l_{mid} = 18$ for all experiments.

D.5.4 Impact of Fusion Strategy

We evaluate the sensitivity of feature fusion in Table 11 by comparing the Concatenation ($\mathbf{F}_{\text{cat}} = [\mathbf{F}^{(l_{\text{mid}})} \parallel \mathbf{F}^{(L)}]$) with a Cross-Attention mechanism:

$$\mathbf{F}_{\text{attn}} = \mathbf{F}^{(l_{\text{mid}})} + \text{CrossAttn}(\mathbf{F}^{(L)}, \mathbf{F}^{(l_{\text{mid}})}, \mathbf{F}^{(l_{\text{mid}})}), \tag{38}$$

where $\mathbf{F}^{(L)}$ serves as the query to extract details from $\mathbf{F}^{(l_{\text{mid}})}$. The two strategies differ by only 0.9%, confirming that our method is insensitive to classifier design. We therefore adopt the more efficient Concatenation for all experiments.

D.6 Robustness Analysis

To comprehensively assess the stability of our detector against common image degradations, we introduce four types of random perturbations: Gaussian Blur, Block-wise Distortion, Contrast Change, and JPEG Compression. We quantify the impact of these disturbances using the Relative AUC Variation (R_{AUC}), defined as:

$$R_{\text{AUC}} = \frac{\text{AUC}_{\text{ptb}} - \text{AUC}_{\text{raw}}}{\text{AUC}_{\text{raw}}}, \tag{39}$$

Table 9: Sensitivity analysis of the mask threshold τ on classical forensic benchmarks (Frame-Level AUC (%)). Best results are in **bold**.

τ	CDFv1	CDFv2	DFDCP	DFDC	DFD	Avg.
0	88.6	88.5	89.1	84.1	91.2	88.3
0.1	**90.9**	**89.1**	88.8	**84.5**	**91.3**	**88.9**
0.5	89.1	88.7	**89.3**	84.3	**91.3**	88.5
0.9	88.9	88.7	89.1	84.2	91.1	88.4
1.3	87.9	86.4	85.3	82.3	90.1	86.4

Table 10: Ablation study on the selection of intermediate layers for feature fusion. The index indicates the encoder layer used for extracting shallow features. Best results are in **bold**.

Layer	CDFv1	CDFv2	DFDCP	DFDC	DFD	Avg.
10	86.6	86.8	85.3	84.1	91.5	86.9
16	88.4	88.2	87.8	84.4	**91.9**	88.1
18	**90.9**	**89.1**	**88.4**	84.5	91.5	**88.9**
20	90.3	88.9	85.5	**85.0**	91.2	88.2

where AUC_{raw} and AUC_{ptb} denote performance on the original and perturbed datasets, respectively. R_{AUC} values closer to zero indicate higher robustness.

As shown in Figure 10, all methods degrade as perturbation severity increases, but CoRIT exhibits the slowest decay rate on average, particularly under heavy JPEG Compression (levels 1–5), which confirms the effectiveness of our noise filtering mechanism. One exception is Block-wise Distortion, where CoRIT shows higher sensitivity. We attribute this to CLIP's intrinsic prior on semantic integrity. Unlike local noise, block-wise occlusion disrupts the semantic integrity of the image. Masking key regions introduces ambiguity that itself resembles a form of manipulation. We therefore consider this sensitivity reasonable behavior for deepfake detectors.

Table 11: Ablation study on fusion architecture. We compare feature concatenation (Concat) with a cross-attention mechanism (CrossAttn). Best results are in **bold**.

Type	CDFv1	CDFv2	DFDCP	DFDC	DFD	Avg.
CrossAttn	**91.5**	88.4	85.6	83.9	90.4	88.0
Concat	90.9	**89.1**	**88.4**	**84.5**	**91.5**	**88.9**

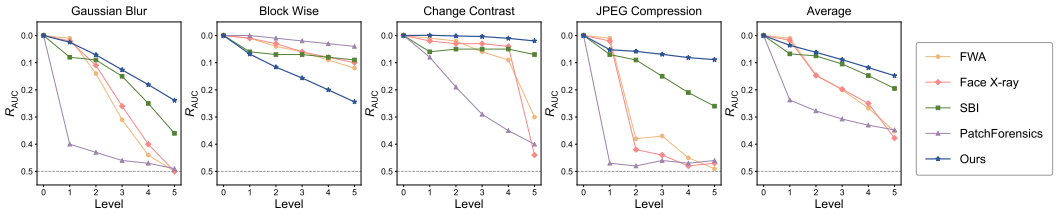


Figure 10: Robustness evaluation under different perturbation levels: Gaussian Blur, Block-wise Distortion, Contrast Change, and JPEG Compression. Relative AUC Variation curves show AUC degradation as perturbation severity increases.

E Additional Evidence for the Fidelity–Generalization Link

E.1 Fidelity and Semantic Displacement

E.1.1 FID Protocol

For each FF++ manipulation, we compute FID between the real-frame distribution and its forged-frame distribution. We report both the test split and the combined training and test sets because FID depends on sample count; the two settings differ in magnitude but give the same ordering.

Table 12: Real–fake FID for the four FF++ manipulations. Each count gives the number of real and forged frames per manipulation.

Sampling range	Frames	NT	F2F	FS	DF
Test	$1{,}120 + 1{,}120$	8.808	11.194	13.403	16.451
Train + test	$6{,}871 + 6{,}871$	2.623	3.055	4.099	5.396

E.1.2 Text-Semantic Subspace

We use $4{,}480$ matched real–forged test-frame pairs, with $1{,}120$ pairs from each manipulation. The frozen CLIP-ViT-L/14 image encoder produces an L2-normalized feature $\mathbf{v}_i \in \mathbb{R}^{768}$ for each frame, and the paired displacement is

$$\Delta \mathbf{v}_i = \mathbf{v}_i^{\text{fake}} - \mathbf{v}_i^{\text{real}}. \tag{40}$$

The text-semantic basis is built from 34 positive–negative concepts covering expression, pose, facial attributes, actions, and scene relations. Each concept is instantiated with five prompt templates; the normalized positive–negative CLIP text-feature differences are averaged, and SVD retains the 19 right-singular directions needed to preserve approximately 95% of their energy. Terms that directly name authenticity, manipulation methods, artifacts, blur, noise, compression, or frequency are excluded from the concept set.

Let $\mathbf{Q} \in \mathbb{R}^{768\times 19}$ denote the orthonormal text basis. We compute the semantic displacement and its magnitude as

$$\mathbf{z}_i = \mathbf{Q}^\top \Delta \mathbf{v}_i, \qquad D_{\text{sem},i} = \|\mathbf{z}_i\|_2. \tag{41}$$

The resulting mean semantic distances are reported in Table 13. They follow the same ordering across forgery types as FID, COR, and held-out AUC in the main text, namely DF $>$ FS $>$ F2F $>$

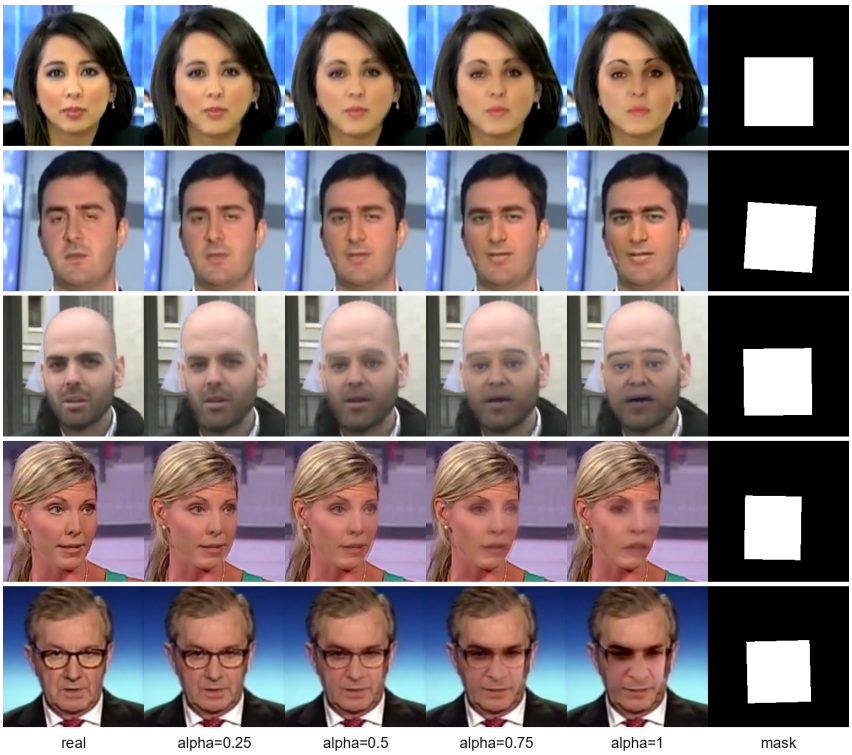


Figure 11: Qualitative examples of the paired within-DF fidelity intervention. Each row shows the matched real frame, interpolated samples, the original DF frame ($\alpha = 1$), and its manipulation mask. Decreasing α attenuates the same pair-specific DF residual, producing a smooth transition toward the real-frame endpoint while keeping the source manipulation and spatial support fixed.

NT. This pattern suggests that higher-fidelity forgeries are associated with lower COR, smaller semantic displacement, and poorer detector generalization.

Table 13: Mean 19-dimensional semantic displacement on matched test-frame pairs.

Manipulation	NT	F2F	FS	DF
Mean D_{sem}	0.046	0.062	0.103	0.115

E.2 Controlled Fidelity Intervention

For each matched real–DF pair, the forged frame is replaced by

$$\mathbf{x}_\alpha = \mathbf{x}_{\text{real}} + \alpha(\mathbf{x}_{\text{fake}} - \mathbf{x}_{\text{real}}), \qquad \alpha \in (0, 1]. \tag{42}$$

This construction preserves the source manipulation and the spatial pattern of each pair-specific DF residual while scaling its magnitude by α. Reducing α moves the constructed sample toward its matched real frame and serves as an operational proxy for increasing fidelity. As illustrated in Figure 11, the interpolation yields a visually smooth progression while largely preserving the frame correspondence, scene context, and spatial manipulation support. For each α, we repeat the SAM radius sweep to locate the coarse COR interval and choose ρ on the validation split.

F Theoretical Derivations and Supporting Analysis

This appendix collects the complete theoretical framework, detailed mathematical derivations, and supporting analyses.

F.1 Decomposition of the Hessian Matrix

Let $\mathbf{x}$ be the input data drawn from the true data distribution $q(\mathbf{x})$, and let $p_{\mathbf{w}_t}(\mathbf{x})$ denote the parameterized model distribution at optimization step t. Under the standard classification setting, the objective function $\mathcal{L}(\mathbf{w}_t)$ is the expected Negative Log-Likelihood (NLL) loss:

$$\mathcal{L}_t = \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[-\log p_{\mathbf{w}_t}(\mathbf{x})]. \tag{43}$$

The stochastic gradient for a single sample $\mathbf{x}$, denoted by $\mathbf{g}(\mathbf{x}, \mathbf{w}_t)$, is computed as:

$$\mathbf{g}(\mathbf{x}, \mathbf{w}_t) = -\nabla_{\mathbf{w}_t} \log p_{\mathbf{w}_t}(\mathbf{x}) = -\frac{\nabla_{\mathbf{w}_t} p_{\mathbf{w}_t}(\mathbf{x})}{p_{\mathbf{w}_t}(\mathbf{x})}. \tag{44}$$

The expected global gradient is therefore $\nabla\mathcal{L}_t = \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\mathbf{g}(\mathbf{x}, \mathbf{w}_t)]$.

Next, we compute the Hessian matrix for a single sample by differentiating the gradient in Eq. (44) with respect to $\mathbf{w}_t$. Applying the quotient rule yields:

$$\begin{aligned}
\mathbf{H}(\mathbf{x}, \mathbf{w}_t) &= -\nabla^2_{\mathbf{w}_t} \log p_{\mathbf{w}_t}(\mathbf{x}) \\
&= -\nabla_{\mathbf{w}_t} \left(\frac{\nabla_{\mathbf{w}_t} p_{\mathbf{w}_t}(\mathbf{x})}{p_{\mathbf{w}_t}(\mathbf{x})} \right) \\
&= -\frac{\nabla^2_{\mathbf{w}_t} p_{\mathbf{w}_t}(\mathbf{x}) \cdot p_{\mathbf{w}_t}(\mathbf{x}) - \nabla_{\mathbf{w}_t} p_{\mathbf{w}_t}(\mathbf{x}) \nabla_{\mathbf{w}_t} p_{\mathbf{w}_t}(\mathbf{x})^\top}{(p_{\mathbf{w}_t}(\mathbf{x}))^2} \\
&= \left(\frac{\nabla_{\mathbf{w}_t} p_{\mathbf{w}_t}(\mathbf{x})}{p_{\mathbf{w}_t}(\mathbf{x})} \right) \left(\frac{\nabla_{\mathbf{w}_t} p_{\mathbf{w}_t}(\mathbf{x})}{p_{\mathbf{w}_t}(\mathbf{x})} \right)^\top - \frac{\nabla^2_{\mathbf{w}_t} p_{\mathbf{w}_t}(\mathbf{x})}{p_{\mathbf{w}_t}(\mathbf{x})} \\
&= \mathbf{g}(\mathbf{x}, \mathbf{w}_t) \mathbf{g}(\mathbf{x}, \mathbf{w}_t)^\top - \frac{\nabla^2_{\mathbf{w}_t} p_{\mathbf{w}_t}(\mathbf{x})}{p_{\mathbf{w}_t}(\mathbf{x})}.
\end{aligned} \tag{45}$$

Taking the expectation over the true data distribution $q(\mathbf{x})$, the expected Hessian matrix $\mathbf{H}_t$ is formulated as:

$$\mathbf{H}_t = \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\mathbf{g}(\mathbf{x}, \mathbf{w}_t)\mathbf{g}(\mathbf{x}, \mathbf{w}_t)^\top] - \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})} \left[\frac{\nabla^2_{\mathbf{w}_t} p_{\mathbf{w}_t}(\mathbf{x})}{p_{\mathbf{w}_t}(\mathbf{x})} \right]. \tag{46}$$

Recall that the uncentered second moment of the gradient relates to the gradient covariance matrix $\mathrm{Cov}[\mathbf{g}_t]$ and the expected gradient $\nabla\mathcal{L}_t$ via the standard identity:

$$\mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})}[\mathbf{g}(\mathbf{x}, \mathbf{w}_t)\mathbf{g}(\mathbf{x}, \mathbf{w}_t)^\top] = \mathrm{Cov}[\mathbf{g}_t] + \nabla\mathcal{L}_t \nabla\mathcal{L}_t^\top. \tag{47}$$

Substituting Eq. (47) into Eq. (46), we obtain the precise structural decomposition of the Hessian:

$$\mathbf{H}_t = \mathrm{Cov}[\mathbf{g}_t] + \nabla\mathcal{L}_t \nabla\mathcal{L}_t^\top - \mathbf{\Xi}_t, \tag{48}$$

where $\mathbf{\Xi}_t = \mathbb{E}_{\mathbf{x}\sim q(\mathbf{x})} \left[\frac{\nabla^2_{\mathbf{w}_t} p_{\mathbf{w}_t}(\mathbf{x})}{p_{\mathbf{w}_t}(\mathbf{x})} \right]$ defines the residual matrix capturing the model misspecification.

F.2 Theoretical Justification for Local Curvature Assumption

In our theoretical framework, the local smoothness constant L_t within the ρ-neighborhood is bounded using the maximum eigenvalue of the Hessian at the center, i.e., $L_t \approx \lambda_{\max}(\mathbf{H}_t)$. This formulation is rigorously grounded in recent theoretical treatments of Sharpness-Aware Minimization (SAM).

Following prior theoretical analysis (Long & Bartlett, 2024), the loss landscape within a local neighborhood can be characterized by a local quadratic approximation. For any weight vector $\mathbf{w}$ near $\mathbf{w}_t$, the loss is approximated as:

$$\mathcal{L}(\mathbf{w}) \approx \mathcal{L}(\mathbf{w}_t) + \nabla\mathcal{L}(\mathbf{w}_t)^\top(\mathbf{w} - \mathbf{w}_t) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_t)^\top \mathbf{H}_t (\mathbf{w} - \mathbf{w}_t), \tag{49}$$

where $\mathbf{H}_t = \nabla^2\mathcal{L}(\mathbf{w}_t)$. The local optimization dynamics and the bounds for the edge of stability in SAM are fundamentally governed by the operator norm of the Hessian, $\|\mathbf{H}_t\|_{\mathrm{op}}$ (Long & Bartlett,

2024). Since the operator norm equals the largest eigenvalue in absolute value, i.e., $\|\mathbf{H}_t\|_{\text{op}} = \max\{|\lambda_{\max}(\mathbf{H}_t)|, |\lambda_{\min}(\mathbf{H}_t)|\}$, this approximation holds when the magnitude of the most negative eigenvalue is dominated by $\lambda_{\max}(\mathbf{H}_t)$, a condition broadly observed in practice after the initial training phase (Sagun et al., 2017; Ghorbani et al., 2019).

$$\|\mathbf{H}_t\|_{\text{op}} \approx \lambda_{\max}(\mathbf{H}_t). \tag{50}$$

To mathematically justify extending this pointwise metric to the supremum over the entire neighborhood $\mathcal{N}(\mathbf{w}_t, \rho)$, we rely on an established framework (Bartlett et al., 2023). By assuming that the third-order derivative tensor $D^3\mathcal{L}$ is B-Lipschitz continuous with respect to the operator norm, the variation of the Hessian along any perturbation vector $\boldsymbol{\epsilon}$ (with $\|\boldsymbol{\epsilon}\|_2 \leq \rho$) can be evaluated via the fundamental theorem of calculus:

$$\nabla^2\mathcal{L}(\mathbf{w}_t + \boldsymbol{\epsilon}) = \mathbf{H}_t + \int_0^1 D^3\mathcal{L}(\mathbf{w}_t + x\boldsymbol{\epsilon})(\boldsymbol{\epsilon}, \cdot, \cdot)\, dx. \tag{51}$$

Applying the B-Lipschitz property bounds the spectral norm for any perturbed weight along this path (Bartlett et al., 2023):

$$\|\nabla^2\mathcal{L}(\mathbf{w}_t + \boldsymbol{\epsilon})\|_{\text{op}} \leq \lambda_{\max}(\mathbf{H}_t) + \mathcal{O}(\rho). \tag{52}$$

Under the Optimization Collapse phenomenon, where ρ is sufficiently small, the error term $\mathcal{O}(\rho)$ becomes vanishingly small. Therefore, taking the supremum over the neighborhood yields:

$$L_t = \sup_{\|\boldsymbol{\epsilon}\|_2 \leq \rho} \|\nabla^2\mathcal{L}(\mathbf{w}_t + \boldsymbol{\epsilon})\|_{\text{op}} \approx \lambda_{\max}(\mathbf{H}_t). \tag{53}$$

This derivation guarantees that $\lambda_{\max}(\mathbf{H}_t)$ serves as a theoretically sound and tight proxy for the local smoothness constant L_t of the population loss. The extension to mini-batch gradients follows naturally: if each per-sample loss $\ell_i(\mathbf{w})$ satisfies L_t-smoothness within $\mathcal{N}(\mathbf{w}_t, \rho)$, then the mini-batch loss $\mathcal{L}_{\mathcal{B}_t} = \frac{1}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t} \ell_i$ inherits the same constant, since for any $\|\boldsymbol{\epsilon}\| \leq \rho$:

$$\begin{aligned} &\|\nabla\mathcal{L}_{\mathcal{B}_t}(\mathbf{w}_t + \boldsymbol{\epsilon}) - \nabla\mathcal{L}_{\mathcal{B}_t}(\mathbf{w}_t)\| \\ &\leq \frac{1}{|\mathcal{B}_t|}\sum_{i\in\mathcal{B}_t} \|\nabla\ell_i(\mathbf{w}_t + \boldsymbol{\epsilon}) - \nabla\ell_i(\mathbf{w}_t)\| \\ &\leq L_t\|\boldsymbol{\epsilon}\|. \end{aligned} \tag{54}$$

This per-sample smoothness assumption is standard in SAM analysis (Foret et al., 2021; Bartlett et al., 2023; Long & Bartlett, 2024).

G Visualization of Region Refinement Mask Across Layers

Figures 12, 13, and 14 provide a comprehensive layer-by-layer visualization of the proposed Region Refinement Mask (RRM) mechanism applied to representative input samples. The visualization reveals two key observations that validate the effectiveness of RRM.

Effective Filtering of Outlier Tokens. The RRM consistently suppresses spurious, semantically irrelevant tokens across all layers. As shown in the per-region maps, tokens that do not conform to the spatial and semantic structure of each region are progressively masked out, producing increasingly clean and compact region representations. This confirms that RRM successfully identifies and removes outlier Tokens that would otherwise introduce noise into the feature aggregation process, thereby improving the overall discriminability of the learned representations.

Layer-wise Distribution of Outlier Tokens. A closer inspection of the layer-wise progression reveals a non-uniform distribution of outlier Tokens across network depth. Specifically, a notably higher proportion of tokens are filtered in the *shallow layers* (*Layer 1–Layer 4*), where low-level features have not yet been fully contextualized, and in the *final layers* (*Layer 21–Layer 24*), where task-specific representations are being consolidated and certain tokens may drift away from the dominant semantic region. In contrast, the intermediate layers (*Layer 5–Layer 20*) exhibit relatively stable region masks, suggesting that the mid-network features are more coherent and require less aggressive filtering. This layer-wise pattern is consistent with the COR evolution in Figure 1(b) of the main text, where intermediate layers exhibit higher COR values, and provides empirical support for the choice of the intermediate fusion layer l_{mid} in the Hierarchical Representation Integration.

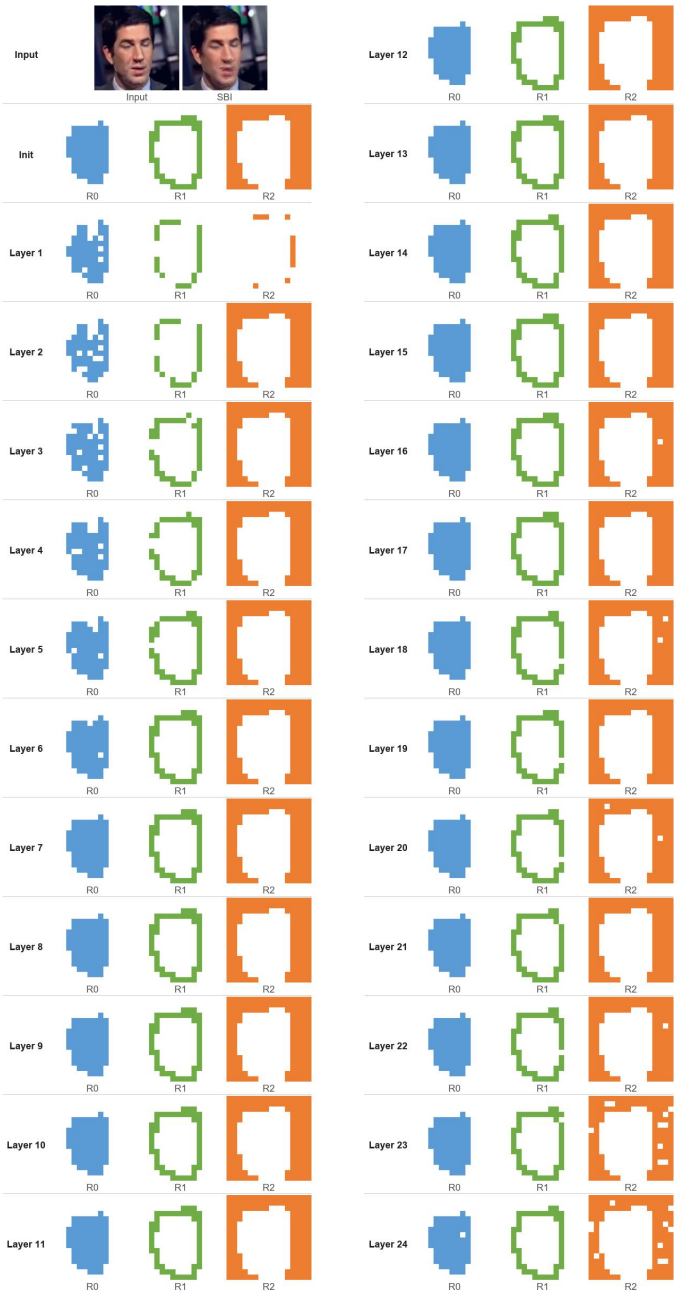


Figure 12: Layer-wise visualization of the Region Refinement Mask (RRM) for sample 1. Each row displays the predicted mask components (Region 0, Region 1, and Region 2) at a specific network layer, accompanied by the input image and the SBI baseline for reference. The progression is visualized from the initialization stage (*Init*) through all 24 transformer layers (*Layers 1 to 24*). Within each map, higher intensity areas indicate tokens retained after refinement, whereas darker regions correspond to suppressed outlier tokens.

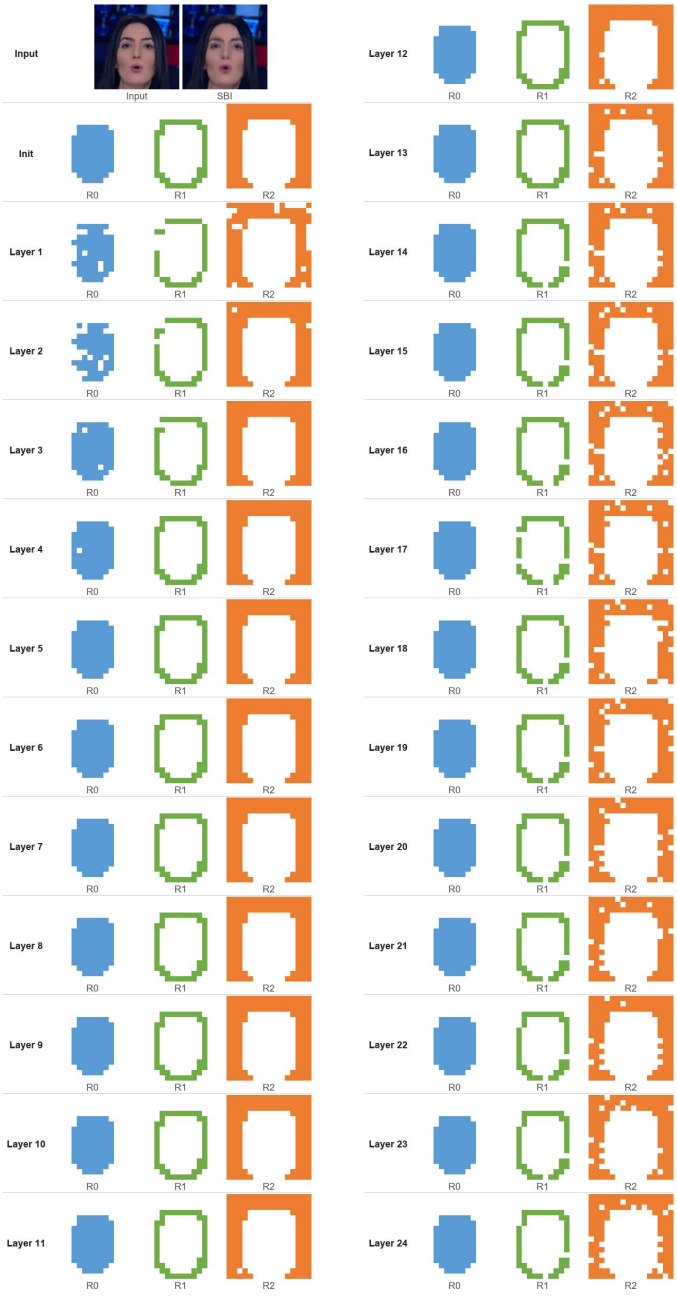


Figure 13: Layer-wise visualization of RRM for sample 2. The layout and color coding follow Figure 12.

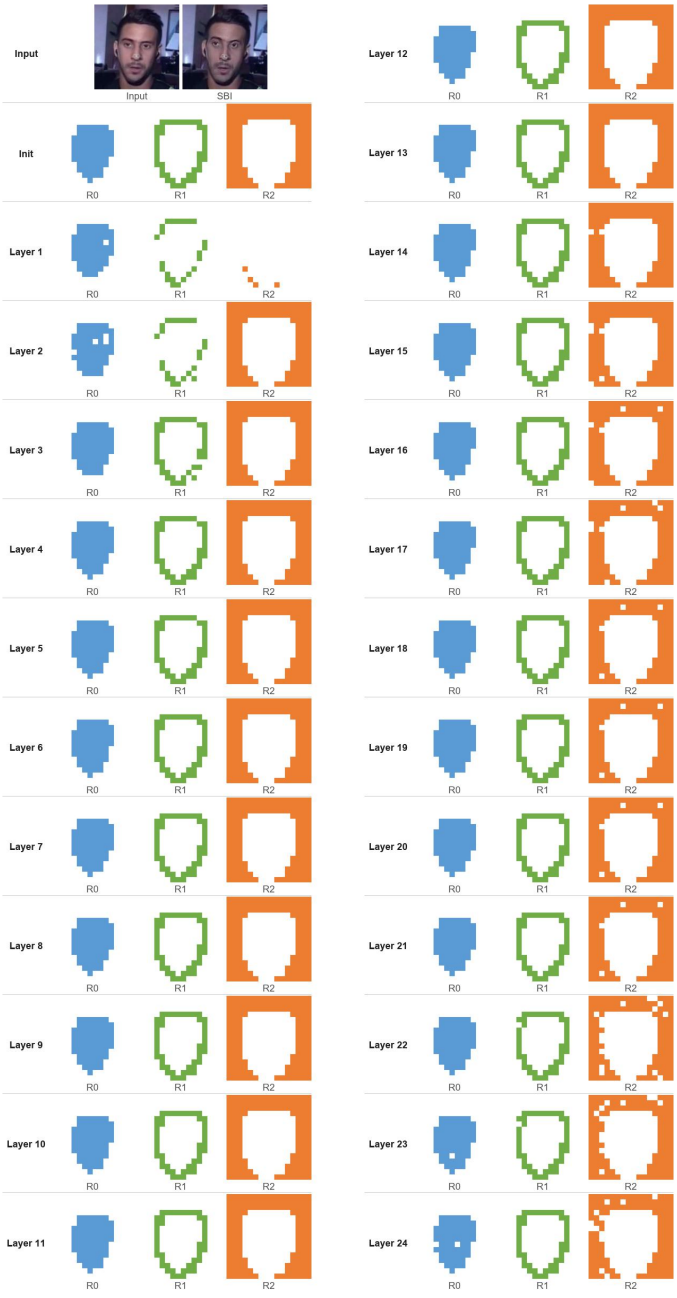


Figure 14: Layer-wise visualization of RRM for sample 3. The layout and color coding follow Figure 12.