SOURCE-GUIDED INFERENCE-TIME RECTIFICATION FOR MORE GENERALIZABLE AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Rapid advances in AI-generated imagery (AIGI) call for reliable and generalizable detectors capable of identifying increasingly diverse and realistic synthetic imagery. However, existing detectors often suffer substantial performance degradation when confronted with unseen generators, manipulation techniques, or post-processing operations. We revisit this problem from the perspective of internal representations and observe a consistent association between detection failures and test representations deviating from reliable source-domain regions, while such deviations are partially rectifiable at inference time. Motivated by these observations, we propose **Source-Guided Subspace Rectification (SGSR)**, a plug-and-play inference-time representation rectification framework that can be integrated into existing AIGI detectors to improve their cross-domain generalization. SGSR models an observed test representation as a perturbed observation of an underlying reliable representation and rectifies it according to a reliable representation distribution estimated from correctly predicted source-domain samples, whose covariance structure enables direction-dependent rectification. SGSR requires no target statistics, pseudo-labels, cross-sample memory, or parameter updates during inference, while introducing only marginal additional computational and storage overhead. Extensive cross-dataset and cross-generator experiments on both Deepfake and synthetic image detection benchmarks demonstrate consistent improvements, with average AUC gains of 0.80 and 4.08 percentage points across 19 Deepfake detectors and 13 synthetic image detectors, respectively.

1 INTRODUCTION

The rapid evolution of generative models, particularly generative adversarial networks (GANs) (Karras et al., 2020) and diffusion models (Rombach et al., 2022), has substantially improved the visual fidelity of synthetic images while lowering the cost and expertise required for their creation. However, these advances also increase the risk of malicious use. Deepfakes, including face swapping (Korshunov & Marcel, 2018) and face reenactment (Thies et al., 2016), can violate personal privacy and facilitate the spread of misinformation (Yan et al., 2025), while the widespread dissemination of synthetic imagery may further erode public trust in digital media (Yan et al., 2026). Therefore, reliable and robust detectors are urgently needed to verify digital-image authenticity.

Existing AIGI detectors have made substantial progress in detection performance, yet still suffer from performance degradation when confronted with unseen generators, manipulation techniques, or post-processing operations (Ojha et al., 2023; Yan et al., 2025). Existing approaches primarily enhance cross-domain generalization during training, while representation shifts arising under unseen test conditions remain largely unaddressed during inference. This limitation motivates us to examine the problem from the perspective of internal representations at inference time.

We revisit generalizable AI-generated image (AIGI)[1] detection from a new perspective: *how do representations behave when the detector fails?* Across four representative detectors—CLIP-LoRA (Hu et al., 2022), BPL (Xu et al., 2026), Effort (Yan et al., 2025), and GenD (Yermakov et al., 2026)—our

[1] In this work, AIGI mainly encompasses both Deepfakes and fully synthetic images.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

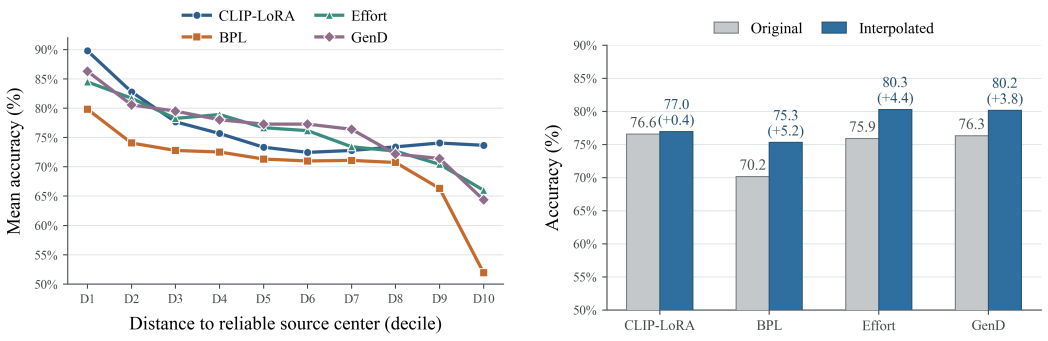


Figure 1: **Preliminary observations for representation rectification.** (a) Detection accuracy decreases as the deviation of test representations from the reliable source center increases. For each layer, test samples are divided into ten Euclidean-distance-based deciles, and accuracy is computed within each decile. (b) Interpolating test representations toward the reliable source center can partially improve detection performance. Further details are provided in Appendix A.1.

preliminary analyses reveal a consistent relationship between representation deviation and prediction correctness. As shown in Fig. 1a, detection accuracy tends to decrease as the distance between test representations and the center of correctly predicted source-domain representations increases. This consistent distance–performance relationship suggests that correctly predicted source representations provide a meaningful characterization of reliable states, motivating our **Modeling Premise 1: reliable representations exhibit a coherent distributional structure that can be estimated from correctly predicted source-domain samples** (Jiang et al., 2018).

This observation motivates a further investigation into whether such representation deviations can be corrected at inference time. For a target representation $\boldsymbol{z}$ (from a test sample), we perform a simple linear interpolation toward the reliable source-domain center $\boldsymbol{\mu}$, estimated from correctly predicted source representations: $\boldsymbol{z}' = \boldsymbol{z} + \lambda(\boldsymbol{\mu} - \boldsymbol{z})$. As shown in Fig. 1b, this simple intervention consistently improves detection performance across all four representative detectors without updating the detector. The partial recovery obtained by moving an observed target representation toward the source-characterized reliable region suggests that its deviation can be viewed as a perturbation from a more reliable latent state. This observation motivates our **Modeling Premise 2: target representations can be modeled as perturbed observations of underlying reliable representations** (Tipping & Bishop, 1999; Singh et al., 2024; Zhou et al., 2024).

Based on these observations, we propose **Source-Guided Subspace Rectification (SGSR)**, a simple yet effective inference-time representation rectification framework for more generalizable AIGI detection. Instead of adapting the detector to an unknown target domain, SGSR uses correctly predicted source-domain representations to characterize a reliable representation distribution and treats each observed test representation as a perturbed observation of an underlying reliable representation. Based on this formulation, the latent reliable representation is inferred in closed form by jointly considering the observed test representation and source-estimated reliable statistics, enabling direction-dependent rectification through the anisotropic covariance structure of the reliable representation space. To ensure stable and efficient estimation in high-dimensional representation spaces, we further approximate the reliable covariance using a low-rank principal subspace. This yields a lightweight sample-wise rectification that leaves the detector parameters unchanged. Moreover, SGSR can be readily applied to both CNN- and Transformer-based detectors and, unlike target-driven test-time adaptation, uses only fixed source-estimated statistics and the current test sample, without estimating target-domain statistics or updating model parameters.

Our main contributions are summarized as follows:

- We empirically identify a consistent association between detection failures and test representations departing from reliable source-domain regions across representative AIGI detectors, and show that such shifts are partially rectifiable at inference time.

- We introduce Source-Guided Subspace Rectification (SGSR), which models a test representation as a perturbed observation of its underlying reliable state and recovers it through covariance-aware low-rank subspace rectification. All required statistics are estimated once from the source domain and fixed thereafter, enabling sample-wise inference without target statistics, pseudo-labels, cross-sample memory, or parameter updates.
- Extensive cross-dataset and cross-generator evaluations on Deepfake and synthetic image detection demonstrate consistent improvements across CNN- and Transformer-based detectors. Further analyses reveal that SGSR induces stronger rectification on incorrectly predicted and more strongly shifted representations, while introducing only marginal storage and inference overhead.

2 Related Work

Generalizable Deepfake Detection. Generalizable deepfake detection aims to learn robust forensic representations that remain discriminative under domain shifts. Earlier methods mainly sought transferable forgery patterns through frequency transformations (Qian et al., 2020; Liu et al., 2021), blending-based synthesis (Li et al., 2020a; Shiohara & Yamasaki, 2022), content/identity disentanglement (Yan et al., 2023a) and latent-space augmentation (Yan et al., 2024). With the emergence of large-scale pretrained models such as CLIP (Radford et al., 2021), recent studies increasingly exploit their rich transferable priors. BPL captures pair-wise discrepancies between real and fake samples in the CLIP representation space (Xu et al., 2026); Effort separates pretrained knowledge and forgery-related representations through orthogonal subspace modeling (Yan et al., 2025); DFF-Adapter introduces fine-grained adapters and dual supervision to jointly learn authenticity and manipulation-specific cues (Zhang et al., 2026); and FSFM learns face-specific transferable representations through self-supervised pretraining on real faces (Wang et al., 2025).

Generalizable Synthetic Image Detection. With the rapid development of generative technologies, synthetic images have proliferated beyond facial content to diverse scenes. Earlier methods mainly sought transferable generation artifacts from data augmentations (Wang et al., 2020), frequency irregularities (Jeong et al., 2022), gradient patterns (Tan et al., 2023), neighboring-pixel dependencies (Tan et al., 2024), and reconstruction discrepancies (Wang et al., 2023). With the emergence of large-scale pretrained models, another line of work leverages their rich transferable knowledge for synthetic image detection (Liu et al., 2024; Zhang et al., 2025). UniFD demonstrates the strong cross-generator transferability of frozen CLIP representations (Ojha et al., 2023); IAPL introduces image-adaptive prompts to extract forgery cues conditioned on individual images (Li et al., 2026); and DGS-Net preserves beneficial pretrained priors while suppressing task-irrelevant information through gradient-space decomposition (Yan et al., 2026). Existing AIGI detection methods predominantly focus on training generalizable classifiers, while little attention has been paid to recovering shifted representations at inference time.

Inference-Time Representation Intervention. Inference-time representation intervention modifies internal representations after training without estimating target-domain statistics or updating model parameters. Early approaches perform sample-wise optimization: PPLM steers language-model hidden states using attribute gradients (Dathathri et al., 2020), while inference-time label-preserving projection maps a test representation onto a source representation manifold before classification (Pandey et al., 2021). Later studies employ more efficient activation manipulation. ITI (Li et al., 2023) and Representation Engineering (Zou et al., 2023) extract behavior-related directions and intervene on hidden activations. For visual tasks, ReAct (Sun et al., 2021) and BATS (Zhu et al., 2022) rectify test-time activations or representations to improve out-of-distribution detection; Robust Feature Inference (Singh et al., 2024) projects representations onto robust spectral subspaces to enhance adversarial robustness; and Feature Clipping (Tao et al., 2025) modifies test representations to improve uncertainty calibration. Recent studies have mainly focused on mitigating hallucinations in multimodal large language models through representation intervention (Yang et al., 2025; Zhuang et al., 2026). These works demonstrate the effectiveness of post-training representation manipulation, but mainly focus on controllable generation, calibration, OOD detection, robustness, or hallucination. We further investigate source-guided inference-time representation intervention for generalizable AI-generated image detection.

3 METHOD

3.1 PROBLEM FORMULATION

Let a trained binary forgery detector be decomposed at layer ℓ as

$$f(\boldsymbol{x}) = h_\ell(g_\ell(\boldsymbol{x}))\,, \tag{1}$$

where g_ℓ maps image $\boldsymbol{x}$ to an intermediate representation $\boldsymbol{z}_\ell \in \mathbb{R}^{d_\ell}$, and h_ℓ denotes the remaining computation and classifier. For a vision transformer, $\boldsymbol{z}_\ell$ is the CLS token at block ℓ; for a CNN, it denotes the globally pooled feature vector. We seek a rectified representation $\boldsymbol{u}_\ell$ that improves the prediction under domain shift while keeping all detector parameters frozen.

Given the labeled source training set $\mathcal{D}_\mathrm{s} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, we collect the layer-$\ell$ representations of samples that are correctly predicted by the trained detector:

$$\mathcal{H}_\ell = \{\boldsymbol{z}_{i,\ell} = g_\ell(\boldsymbol{x}_i) \mid \hat{y}_i = y_i,\ (\boldsymbol{x}_i, y_i) \in \mathcal{D}_\mathrm{s}\}\,. \tag{2}$$

The correctly predicted source representations $\mathcal{H}_\ell$ are used to estimate the detector's reliable representation distribution at layer ℓ. Let Θ_ℓ denote the layer-specific rectification statistics characterizing this reliable representation distribution, estimated from $\mathcal{H}_\ell$.

For a test image $\boldsymbol{x}^t$ from an unseen domain, the detector produces $\boldsymbol{z}_\ell^t = g_\ell(\boldsymbol{x}^t)$. When exposed to unseen test-time conditions, the observed representation may deviate from its latent reliable state, potentially leading to incorrect predictions. Given the observed test representation $\boldsymbol{z}_\ell^t$ and the estimated reliable representation distribution Θ_ℓ, its rectified representation is computed as:

$$\hat{\boldsymbol{u}}_\ell^t = \mathcal{R}\left(\boldsymbol{z}_\ell^t; \Theta_\ell\right), \tag{3}$$

where $\mathcal{R}$ denotes an inference-time rectification operator. The rectified representation is subsequently passed through the remaining detector as $h_\ell(\hat{\boldsymbol{u}}_\ell^t)$. Specifically, the rectification uses only the fixed source-estimated statistics and the current test sample, without requiring target-domain labels, target-domain statistics, or parameter updates.

3.2 PROBABILISTIC REPRESENTATION RECTIFICATION

Building on the two modeling premises introduced in Section 1, we formulate representation rectification from a probabilistic perspective. Let $\boldsymbol{u} \in \mathbb{R}^d$ denote the latent reliable representation and $\boldsymbol{z} \in \mathbb{R}^d$ the observed target representation at a fixed layer (we omit ℓ for readability). Specifically, we model the observed test representation as a perturbed observation of its latent reliable state, while characterizing the latent reliable representation with a source-estimated prior.

Observation model. Given an observed test representation $\boldsymbol{z}$, we regard it as a perturbed observation of an underlying reliable representation $\boldsymbol{u}$. Following standard probabilistic latent-variable formulations, where residual observation uncertainty is commonly modeled using spherical Gaussian noise (Roweis, 1998; Tipping & Bishop, 1999; Lawrence, 2004), we model representation perturbations as additive isotropic Gaussian noise, providing a direction-agnostic characterization of unknown test-time deviations:

$$\boldsymbol{z} = \boldsymbol{u} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \tau^2 \boldsymbol{I}), \tag{4}$$

where τ^2 parameterizes the variance of representation perturbation. This observation model gives

$$p(\boldsymbol{z} \mid \boldsymbol{u}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{u}, \tau^2 \boldsymbol{I}). \tag{5}$$

Reliable-representation prior. Given the correctly predicted source representations, we model the latent distribution of reliable representations with a multivariate Gaussian prior (Lee et al., 2018; Cui et al., 2023). We provide a qualitative examination of this Gaussian approximation in Appendix A.3. Specifically,

$$\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_r), \tag{6}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_r$ are the mean and covariance of the reliable representation distribution, estimated from the correctly predicted source representations $\mathcal{H} = \{\boldsymbol{z}_i\}_{i=1}^n$.

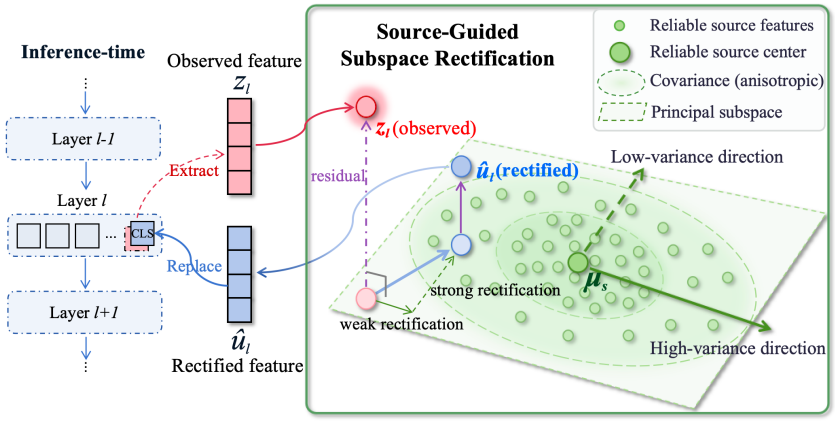


Figure 2: **Overview of Source-Guided Subspace Rectification (SGSR).** SGSR estimates a reliable representation distribution from correctly predicted source-domain representations and rectifies intermediate test representations toward their inferred reliable states. The figure illustrates the ViT implementation using `[CLS]` representations; for CNNs, SGSR instead rectifies global average-pooled (GAP) representations and broadcasts the resulting corrections back to the feature maps.

Given an observed test representation $\boldsymbol{z}$, the latent reliable representation $\boldsymbol{u}$ is treated as an unobserved variable constrained jointly by the observation model and the reliable-representation prior. Bayes' theorem gives the posterior over the latent reliable representation:

$$\begin{aligned} p(\boldsymbol{u} \mid \boldsymbol{z}) &= \frac{p(\boldsymbol{z} \mid \boldsymbol{u})p(\boldsymbol{u})}{p(\boldsymbol{z})} \\ &\propto p(\boldsymbol{z} \mid \boldsymbol{u})p(\boldsymbol{u}) \\ &\propto \exp\left(-\frac{1}{2\tau^2}\|\boldsymbol{z}-\boldsymbol{u}\|_2^2\right)\exp\left[-\frac{1}{2}(\boldsymbol{u}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}_r^{-1}(\boldsymbol{u}-\boldsymbol{\mu})\right], \end{aligned} \tag{7}$$

where $p(\boldsymbol{u})$ denotes the prior distribution of reliable representations.

The latent reliable representation is then inferred via maximum a posteriori (MAP) estimation:

$$\begin{aligned} \hat{\boldsymbol{u}} &= \arg\max_{\boldsymbol{u}} p(\boldsymbol{u} \mid \boldsymbol{z}) = \arg\max_{\boldsymbol{u}} p(\boldsymbol{z} \mid \boldsymbol{u})p(\boldsymbol{u}) \\ &= \arg\min_{\boldsymbol{u}} \left[\frac{1}{2\tau^2}\|\boldsymbol{u}-\boldsymbol{z}\|_2^2 + \frac{1}{2}(\boldsymbol{u}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}_r^{-1}(\boldsymbol{u}-\boldsymbol{\mu})\right]. \end{aligned} \tag{8}$$

The first term preserves fidelity to the observed test representation, whereas the second regularizes the solution toward the reliable source distribution. The resulting quadratic objective can be solved in closed form as:

$$\hat{\boldsymbol{u}} = \boldsymbol{\mu} + \boldsymbol{\Sigma}_r\left(\boldsymbol{\Sigma}_r + \tau^2\mathbf{I}\right)^{-1}(\boldsymbol{z}-\boldsymbol{\mu}). \tag{9}$$

Thus, this closed-form solution rectifies the observed representation according to the covariance structure of the reliable source distribution while retaining its original information. Specifically, directions with higher reliable variance receive weaker rectification, whereas low-variance directions are corrected more strongly.

3.3 Low-Rank Subspace Rectification

Although Eq. 9 provides an exact solution, directly estimating and applying $\boldsymbol{\Sigma}_r \in \mathbb{R}^{d\times d}$ can be unreliable and costly in high-dimensional representation spaces. With finite reliable source samples, the eigensystem of the sample covariance can be sensitive to estimation error, with its accuracy depending on the available sample size and spectral separation (Loukas, 2017). Further empirical evidence for this behavior is provided in Appendix A.4. Therefore, we construct a low-rank approximation of the reliable covariance structure.

Let

$$\mathbf{X} = [\boldsymbol{z}_1 - \boldsymbol{\mu}, \ldots, \boldsymbol{z}_n - \boldsymbol{\mu}] \in \mathbb{R}^{d \times n} \tag{10}$$

denote the centered reliable source representations. We perform the standard singular value decomposition (SVD)

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^\top, \tag{11}$$

where the columns of $\mathbf{U}$ represent the principal directions in the representation space. Since $\boldsymbol{\Sigma}_r = \frac{1}{n-1}\mathbf{X}\mathbf{X}^\top$, the left singular vectors in $\mathbf{U}$ are also the eigendirections of $\boldsymbol{\Sigma}_r$, with corresponding variances $\lambda_i = s_i^2/(n-1)$.

To reduce estimation noise and computational cost, we retain only the top-K principal directions $\mathbf{U}_K = [\mathbf{u}_1, \ldots, \mathbf{u}_K]$ and their corresponding variances $\boldsymbol{\Lambda}_K = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$. The remaining low-energy directions are summarized by a shared residual variance

$$\sigma_\perp^2 = \frac{\|\mathbf{X}\|_F^2 - \sum_{i=1}^K s_i^2}{(n-1)(d-K)}. \tag{12}$$

This yields the structured covariance approximation

$$\widetilde{\boldsymbol{\Sigma}}_r = \mathbf{U}_K \boldsymbol{\Lambda}_K \mathbf{U}_K^\top + \sigma_\perp^2 \left(\mathbf{I} - \mathbf{U}_K \mathbf{U}_K^\top\right). \tag{13}$$

For an observed representation $\boldsymbol{z}$, its centered vector can be decomposed into the retained principal component and the orthogonal residual as

$$\boldsymbol{z} - \boldsymbol{\mu} = \mathbf{U}_K \boldsymbol{c} + \boldsymbol{r}, \qquad \boldsymbol{c} = \mathbf{U}_K^\top (\boldsymbol{z} - \boldsymbol{\mu}), \qquad \boldsymbol{r} = \left(\mathbf{I} - \mathbf{U}_K \mathbf{U}_K^\top\right)(\boldsymbol{z} - \boldsymbol{\mu}). \tag{14}$$

Substituting the low-rank covariance approximation in Eq. 13 into the closed-form solution in Eq. 9, the final rectified representation is given by

$$\hat{\boldsymbol{u}} = \boldsymbol{\mu} + \mathbf{U}_K \,\mathrm{diag}\left(\frac{\lambda_i}{\lambda_i + \tau^2}\right) \boldsymbol{c} + \frac{\sigma_\perp^2}{\sigma_\perp^2 + \tau^2} \boldsymbol{r} \tag{15}$$

The resulting low-rank formulation preserves the dominant structure of the reliable representation distribution while reducing sensitivity to noisy estimation in low-variance directions. It also avoids operating on the full covariance matrix during inference, requiring only projections onto $\mathbf{U}_K$. Consequently, the per-layer computational complexity is reduced from $\mathcal{O}(d^2)$ to $\mathcal{O}(dK)$ for $K \ll d$.

3.4 Inference-Time Rectification

For each intermediate layer ℓ, our rectification requires only the source-estimated statistics

$$\Theta_\ell = \left\{\boldsymbol{\mu}_\ell, \mathbf{U}_{K,\ell}, \{\lambda_{\ell,i}\}_{i=1}^K, \sigma_{\perp,\ell}^2\right\}, \tag{16}$$

together with the layer-wise rectification parameter τ_ℓ^2. All layer-wise statistics are estimated once from correctly predicted source representations and fixed thereafter. Rectification is performed independently for each test sample without accessing source samples, estimating target statistics, or updating detector parameters; the complete procedure is summarized in Appendix A.2.

During inference, the detector follows its original forward computation and applies Eq. 15 at each intermediate layer. For a vision transformer, $\boldsymbol{z}_\ell$ corresponds to the `[CLS]` token. The rectified representation $\hat{\boldsymbol{u}}_\ell$ simply replaces the original `[CLS]` token before the token sequence is passed to the next block.

For a CNN, we first obtain a compact channel representation by global average pooling,

$$\boldsymbol{z}_\ell = \mathrm{GAP}(\mathbf{F}_\ell), \tag{17}$$

where $\mathbf{F}_\ell \in \mathbb{R}^{C_\ell \times H_\ell \times W_\ell}$ denotes the intermediate feature map. After rectifying $\boldsymbol{z}_\ell$ into $\hat{\boldsymbol{u}}_\ell$, the resulting correction is broadcast back to the feature map:

$$\widehat{\mathbf{F}}_\ell = \mathbf{F}_\ell + (\hat{\boldsymbol{u}}_\ell - \boldsymbol{z}_\ell) \otimes \mathbf{1}_{H_\ell \times W_\ell}. \tag{18}$$

The updated feature map is then forwarded through the remaining CNN layers.

Rectification is applied sequentially across all intermediate layers in forward order, so that each subsequent representation is computed from the rectified preceding states. Consequently, multi-layer rectification is performed progressively within a single forward pass.

4 EXPERIMENTS

Rectification Setup. For each detector, we extract intermediate representations from the source training set and retain only correctly classified samples. Real and fake samples are balanced before estimating the reliable source statistics.

For each intermediate layer ℓ, PCA is independently performed on the centered representations. Let $\lambda_{\ell,1} \geq \lambda_{\ell,2} \geq \cdots \geq \lambda_{\ell,d_\ell}$ denote the variances along the principal directions. We determine the retained dimension K_ℓ as the minimum number of principal components whose cumulative explained variance reaches η:

$$K_\ell = \min \left\{ K : \frac{\sum_{i=1}^{K} \lambda_{\ell,i}}{\sum_{i=1}^{d_\ell} \lambda_{\ell,i}} \geq \eta \right\}. \tag{19}$$

To ensure consistent rectification across layers with different representation scales, we define the layer-specific perturbation variance relative to the average reliable variance at each layer:

$$\tau_\ell^2 = \alpha\, \bar{\sigma}_\ell^2, \qquad \bar{\sigma}_\ell^2 = \frac{1}{d_\ell} \operatorname{Tr}\left(\mathbf{\Sigma}_{r,\ell}\right). \tag{20}$$

We set $\eta = 0.9$ and $\alpha = 0.2$ throughout all main experiments.

In all subsequent experiments, rectification is applied to all intermediate layers. No detector parameters are updated, and no target-domain labels or statistics are used during inference.

4.1 DEEPFAKE IMAGE DETECTION

Evaluation tasks and datasets. For Deepfake image detection, we follow the cross-dataset evaluation protocol and the preprocessing and training pipeline of DeepfakeBench (Yan et al., 2023b). Specifically, all detectors are trained on the c23 version of FaceForensics++ (FF++) (Rossler et al., 2019), which serves as the source domain. Cross-dataset generalization is evaluated on five unseen datasets: DeepFakeDetection (DFD) (Dufour et al., 2019), Celeb-DF-v2 (CDF-v2) (Li et al., 2020b), Deepfake Detection Challenge Preview (DFDCP) (Dolhansky et al., 2019), Deepfake Detection Challenge (DFDC) (Dolhansky et al., 2020), and WildDeepfake (WDF) (Zi et al., 2020).

Evaluated detectors. We evaluate 19 detectors spanning three groups. *CNN-based forgery detectors:* Xception (Chollet, 2017), ResNet-50 (He et al., 2016), SBI (Shiohara & Yamasaki, 2022), RECCE (Cao et al., 2022), UCF (Yan et al., 2023a), and LSDA (Yan et al., 2024). *Classical fine-tuning strategies:* five CLIP ViT-L/14 (Radford et al., 2021) variants using full fine-tuning, LayerNorm tuning (Qi et al., 2022), LoRA (Hu et al., 2022), visual prompt tuning (VPT) (Jia et al., 2022), and BitFit (Ben Zaken et al., 2022), denoted as CLIP-FFT, CLIP-LN, CLIP-LoRA, CLIP-VPT, and CLIP-BitFit, respectively. *Forgery-specific ViT/VFM detectors:* UIA-ViT (Zhuang et al., 2022), BPL (Xu et al., 2026), UDD (Fu et al., 2025), Effort (Yan et al., 2025), DFF-Adapter (Zhang et al., 2026), IAPL (Li et al., 2026), and the CLIP ViT-L/14 and DINOv3 ViT-L/16 variants of GenD (Yermakov et al., 2026), denoted as GenD-CLIP-L and GenD-DINOv3.

Evaluation Analysis. Table 1 reports the cross-dataset results of the original detectors and the improvements obtained after rectification. Overall, SGSR improves the average AUC of all 19 detectors, with an average gain of 0.80 percentage points. Among the 95 detector–dataset combinations, 87 (91.6%) achieve positive improvements. Consistent gains are observed across CNN-based detectors, classical CLIP adaptation strategies, and forgery-specific ViT/VFM detectors, with average improvements of 0.388, 1.456, and 0.689 points, respectively. Despite minor degradation in a few cases, all detectors achieve positive average gains across the five unseen datasets.

4.2 SYNTHETIC IMAGE DETECTION

Evaluation tasks and datasets. For synthetic image detection, we follow the UniversalFakeDetect benchmark (Ojha et al., 2023) and the standard four-class ProGAN training setting (Wang et al., 2020; Liu et al., 2024). The evaluated benchmark subsets include ProGAN (Karras et al., 2018),

Table 1: Cross-dataset Deepfake image detection results in terms of frame-level AUC (%). Each entry is reported as *baseline AUC* (ΔAUC). Positive improvements are shown in bold.

Method	DFD	CDF-v2	DFDCP	DFDC	WDF	Avg.
CNN-based forgery detectors						
ResNet-50 (He et al., 2016)	78.160 (**+0.998**)	61.285 (**+0.517**)	63.848 (**+0.058**)	65.413 (**+0.843**)	74.321 (-1.333)	68.606 (**+0.217**)
Xception (Chollet, 2017)	77.875 (**+0.475**)	67.373 (**+0.365**)	66.143 (**+0.606**)	65.373 (**+0.287**)	70.223 (-0.038)	69.397 (**+0.339**)
RECCE (Cao et al., 2022)	85.056 (**+0.078**)	60.609 (**+0.292**)	72.101 (**+0.507**)	67.298 (**+0.095**)	63.120 (**+0.401**)	69.637 (**+0.275**)
SBI (Shiohara & Yamasaki, 2022)	73.684 (**+0.962**)	62.486 (**+0.201**)	65.905 (**+0.514**)	60.227 (**+0.297**)	52.261 (**+0.135**)	62.913 (**+0.422**)
UCF (Yan et al., 2023a)	74.655 (**+0.593**)	65.565 (**+0.460**)	62.906 (**+0.676**)	64.482 (**+0.394**)	64.054 (**+1.128**)	66.333 (**+0.650**)
LSDA (Yan et al., 2024)	75.146 (-0.013)	62.833 (**+0.005**)	57.273 (**+0.584**)	60.785 (**+0.058**)	61.907 (**+1.482**)	63.589 (**+0.423**)
Classical CLIP adaptation strategies						
CLIP-FFT (Radford et al., 2021)	90.472 (**+1.939**)	82.932 (**+1.635**)	80.655 (**+1.478**)	78.658 (**+1.979**)	81.038 (**+2.962**)	82.751 (**+1.996**)
CLIP-LN (Qi et al., 2022)	72.768 (-0.544)	71.719 (-0.317)	70.649 (**+3.007**)	66.467 (**+0.965**)	77.872 (**+2.156**)	71.895 (**+1.054**)
CLIP-LoRA (Hu et al., 2022)	86.256 (**+0.519**)	82.533 (**+0.282**)	81.887 (**+0.678**)	78.255 (**+0.482**)	79.217 (**+0.173**)	81.630 (**+0.427**)
CLIP-BitFit (Ben Zaken et al., 2022)	93.410 (**+0.575**)	89.511 (**+0.953**)	85.544 (**+2.013**)	80.692 (**+0.966**)	83.683 (**+1.820**)	86.568 (**+1.266**)
CLIP-VPT (Jia et al., 2022)	91.192 (**+1.625**)	78.700 (**+2.758**)	78.125 (**+2.558**)	77.564 (**+2.708**)	81.943 (**+3.020**)	81.505 (**+2.534**)
Forgery-specific ViT/VFM detectors						
UIA-ViT (Zhuang et al., 2022)	79.114 (**+0.171**)	75.254 (**+0.321**)	71.334 (**+0.219**)	71.255 (**+0.320**)	70.826 (**+1.717**)	73.557 (**+0.550**)
UDD (Fu et al., 2025)	78.956 (-0.093)	70.533 (**+1.323**)	75.564 (**+0.027**)	74.701 (**+0.000**)	70.052 (**+1.678**)	73.961 (**+0.587**)
Effort (Yan et al., 2025)	93.710 (**+0.368**)	85.622 (**+0.460**)	85.272 (**+0.298**)	81.315 (**+1.127**)	86.123 (**+2.224**)	86.408 (**+0.895**)
DFF-Adapter (Zhang et al., 2026)	92.224 (**+0.446**)	87.958 (**+0.998**)	83.671 (**+0.205**)	82.212 (**+0.740**)	82.301 (**+1.498**)	85.673 (**+0.777**)
GenD-CLIP-L (Yermakov et al., 2026)	94.167 (**+0.503**)	87.803 (**+0.377**)	85.842 (**+0.466**)	82.842 (**+0.510**)	82.545 (**+1.066**)	86.640 (**+0.585**)
GenD-DINOv3 (Yermakov et al., 2026)	90.851 (**+0.294**)	83.667 (**+0.265**)	83.987 (-0.088)	82.023 (**+0.119**)	78.174 (**+0.089**)	83.740 (**+0.136**)
IAPL (Li et al., 2026)	93.087 (**+0.150**)	85.435 (**+0.804**)	84.705 (**+0.074**)	81.772 (**+0.525**)	81.206 (**+1.618**)	85.241 (**+0.634**)
BPL (Xu et al., 2026)	91.347 (**+1.058**)	82.478 (**+1.865**)	81.640 (**+1.825**)	77.895 (**+0.220**)	77.726 (**+1.767**)	82.217 (**+1.347**)

Table 2: Cross-generator synthetic image detection AUC (%), grouped by generation category. Results are reported as baseline AUC (ΔAUC). "Avg." denotes the macro-average over the seven generation categories. Positive improvements are shown in bold.

Method	GAN-based	Deepfakes	Low-level	Guided	LDM	GLIDE	DALL-E	Avg.
Classical adaptation strategies								
CLIP-FFT (Radford et al., 2021)	99.35 (**+0.13**)	87.09 (**+10.69**)	68.87 (**+26.89**)	86.06 (**+10.47**)	96.54 (**+3.39**)	94.88 (**+4.85**)	97.75 (**+2.13**)	90.08 (**+8.36**)
CLIP-LN (Qi et al., 2022)	99.84 (**+0.00**)	94.48 (**+2.74**)	74.88 (**+22.19**)	89.07 (**+9.68**)	98.97 (**+0.99**)	95.88 (**+4.02**)	98.96 (**+0.95**)	93.16 (**+5.80**)
CLIP-LoRA (Hu et al., 2022)	99.57 (**+0.27**)	87.71 (**+6.00**)	64.48 (**+26.69**)	87.14 (**+8.63**)	98.23 (**+1.74**)	88.19 (**+10.35**)	98.74 (**+1.25**)	89.15 (**+7.85**)
CLIP-VPT (Jia et al., 2022)	99.84 (-0.25)	95.84 (**+1.75**)	90.90 (**+1.16**)	89.09 (**+7.24**)	98.42 (**+1.16**)	98.27 (**+1.35**)	99.49 (**+0.34**)	95.98 (**+1.82**)
CLIP-BitFit (Ben Zaken et al., 2022)	99.67 (**+0.17**)	94.77 (**+2.28**)	81.82 (**+12.77**)	86.67 (**+11.05**)	98.62 (**+1.35**)	96.41 (**+3.35**)	98.80 (**+1.14**)	93.82 (**+4.59**)
Task-specific detectors								
CNNSpot (Wang et al., 2020)	83.82 (**+1.52**)	97.35 (-0.19)	90.47 (-0.46)	86.53 (**+0.48**)	94.20 (**+1.61**)	90.43 (**+2.48**)	76.27 (**+3.64**)	88.44 (**+1.30**)
UniFD (Ojha et al., 2023)	99.58 (**+0.06**)	86.94 (**+10.84**)	80.42 (**+11.08**)	92.07 (**+2.13**)	96.92 (**+2.49**)	98.52 (**+0.85**)	98.27 (**+1.14**)	93.25 (**+4.09**)
NPR (Tan et al., 2024)	94.06 (**+0.27**)	70.95 (**+1.64**)	90.57 (**+3.53**)	92.10 (**+0.96**)	99.34 (**+0.12**)	98.72 (**+0.59**)	96.40 (**+0.64**)	91.73 (**+1.11**)
C2P-CLIP (Tan et al., 2025)	99.90 (-0.77)	98.42 (-1.51)	96.95 (**+0.32**)	92.28 (**+6.00**)	99.89 (**+0.06**)	99.85 (-0.08)	99.80 (**+0.05**)	98.15 (**+0.58**)
VIB-Net (Zhang et al., 2025)	98.63 (**+0.18**)	88.89 (**+8.94**)	86.91 (**+6.42**)	82.88 (**+9.83**)	93.39 (**+5.90**)	98.27 (**+1.22**)	95.35 (**+3.93**)	92.05 (**+5.20**)
Effort (Yan et al., 2025)	99.77 (-3.52)	92.81 (**+4.98**)	82.13 (**+7.76**)	90.00 (**+7.71**)	98.48 (**+1.23**)	97.29 (**+2.40**)	99.09 (**+0.59**)	94.22 (**+3.02**)
GenD (Yermakov et al., 2026)	97.14 (**+1.77**)	87.15 (**+9.08**)	89.03 (-0.14)	84.78 (**+7.25**)	94.94 (**+3.90**)	92.19 (**+5.48**)	98.57 (**+1.10**)	91.97 (**+4.06**)
CausalCLIP (Liu et al., 2026)	99.60 (-0.12)	89.89 (**+8.05**)	70.77 (**+20.61**)	87.62 (**+5.79**)	97.90 (**+1.47**)	98.94 (**+0.47**)	98.75 (**+0.61**)	91.92 (**+5.27**)
Avg.	97.75 (-0.02)	90.18 (**+5.02**)	82.17 (**+10.68**)	88.18 (**+6.71**)	97.37 (**+1.95**)	95.99 (**+2.87**)	96.63 (**+1.35**)	92.61 (**+4.08**)

CycleGAN (Zhu et al., 2017), BigGAN (Brock et al., 2019), StyleGAN (Karras et al., 2019), GauGAN (Park et al., 2019), StarGAN (Choi et al., 2018), Deepfakes (Rossler et al., 2019), SITD (Chen et al., 2018), SAN (Dai et al., 2019), Guided Diffusion (Dhariwal & Nichol, 2021), three LDM variants (Rombach et al., 2022), three GLIDE variants (Nichol et al., 2022), and DALL-E (Ramesh et al., 2021). Additional synthetic-image results are provided in Appendix A.9.

Evaluated detectors. We evaluate 13 detectors spanning two groups. The first includes five classical CLIP adaptation strategies: full fine-tuning, LayerNorm tuning, LoRA, VPT, and BitFit. The second includes eight task-specific detectors: CNNSpot (Wang et al., 2020), UniFD (Ojha et al., 2023), NPR (Tan et al., 2024), C2P-CLIP (Tan et al., 2025), VIB-Net (Zhang et al., 2025), Effort (Yan et al., 2025), GenD (Yermakov et al., 2026), and CausalCLIP (Liu et al., 2026).

Evaluation Analysis. As shown in Table 2, SGSR improves the average performance across the 13 evaluated detectors by 4.08 AUC points. In terms of average precision (AP), SGSR also improves the average performance from 93.42 to 96.53, yielding a gain of $+3.10$ points. The gains are particularly pronounced on Deepfakes, low-level synthesis, and Guided Diffusion, while the GAN-based group remains nearly unchanged. One possible factor is that the GAN-based group may be relatively closer to the ProGAN source setting and already exhibits near-saturated baseline

Table 3: Ablation of different rectification formulations averaged over ResNet-50, Effort, and CLIP-LoRA. Baseline reports the average AUC (%), while the remaining rows report ΔAUC (%).

Method	DFD	CDF-v2	DFDCP	DFDC	WDF	Avg.
Baseline	86.042	76.480	77.002	74.994	79.887	78.881
Mean-Interpolation Rectification	-0.038	+0.980	+0.942	+0.339	-1.945	+0.056
Full-Covariance Rectification	+0.398	+0.418	+0.092	+0.409	**+0.690**	+0.401
SGSR (Ours)	**+0.628**	**+0.420**	**+0.345**	**+0.817**	+0.355	**+0.513**

Table 4: Rectification behavior across four representative detectors and five unseen Deepfake datasets.

Group	Representation Variance	Rectification Magnitude	Prediction Change Magnitude	Prediction Transition
Correct (C)	0.01942	0.06064	5.226	$C \to W$: 3.75%
Wrong (W)	0.02470	0.06440	12.014	$W \to C$: 15.67%

performance, leaving limited room for further improvement; additional rectification may even perturb the detector's internal representations and decision process. These results suggest that SGSR can provide larger benefits on more challenging cross-generator settings.

4.3 Ablation Study and Analysis

Ablation on Rectification Formulation. We compare three rectification formulations on ResNet-50, Effort, and CLIP-LoRA: mean-interpolation rectification used in the preliminary analysis in Sec. 1, full-covariance rectification following Eq. 9, and Source-Guided Subspace Rectification (SGSR) following Eq. 15. As shown in Table 3, SGSR achieves the best average improvement, while mean interpolation shows less stable cross-domain behavior and full-covariance modeling provides more consistent but smaller gains.

Rectification Behavior Analysis. We further analyze the rectification behavior of SGSR across four representative detectors: ResNet-50, Effort, GenD-CLIP-L and BPL. As shown in Table 4, wrongly predicted samples exhibit larger representation variance and undergo stronger representation- and prediction-level changes, while the recovery rate (15.67%) is higher than the degradation rate (3.75%). The definitions of the analysis metrics, together with the corresponding analytical derivations and extended results, are provided in Appendix A.10.

Efficiency Analysis. SGSR introduces only marginal overhead: on CLIP-LoRA and ResNet-50, inference latency increases by 0.64% and 3.60%, with storage overheads of 1.08% and 0.12%, respectively. Detailed efficiency results, including comparisons with full-covariance rectification, are provided in Appendix A.5.

5 Conclusion

In this work, we revisit generalizable AI-generated image detection from a representation-space perspective. Our empirical analyses reveal a consistent association between representation deviation and detection failure, and further show that such deviations can be partially rectified without modifying the detector. Based on these observations, we propose Source-Guided Subspace Rectification (SGSR), a plug-and-play framework that models test representations as perturbed observations of underlying reliable representations and rectifies them using source-estimated reliable representation statistics. By combining rectification with low-rank subspace modeling, SGSR provides an efficient and stable mechanism for intervening on intermediate representations without requiring target labels, target statistics, pseudo-labels, cross-sample memory, or parameter updates. Extensive cross-dataset and cross-generator experiments on Deepfake and synthetic image detection demonstrate consistent improvements across diverse detector architectures and adaptation strategies, while incurring limited additional computational and storage overhead. Overall, inference-time representation rectification provides an effective complementary mechanism to existing training-based generalization methods, improving the cross-domain generalization of AIGI detectors under unseen distribution shifts.

AI USE STATEMENT

Generative AI tools were used to assist with writing and language polishing, literature retrieval and discovery, LaTeX formatting, and code debugging. All scientific content, experimental results, analyses, and conclusions were reviewed and verified by the authors. Any AI-assisted outputs were manually checked before being incorporated into the manuscript. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work aims to improve the generalization of detectors for identifying AI-generated and manipulated images, with the goal of supporting digital media authenticity and mitigating the misuse of synthetic content. Our experiments are conducted on established research benchmarks, and we do not collect or release new personal data. The proposed method does not introduce new image-generation or manipulation capabilities. Nevertheless, AI-generated image detectors may still produce false positives or false negatives under unseen conditions. Therefore, such detectors should not be treated as the sole evidence for determining the authenticity or provenance of digital content, especially in high-stakes applications.

REPRODUCIBILITY STATEMENT

We provide the complete formulation and closed-form solution of SGSR in Sec. 3, together with the layer-wise inference procedure in Appendix A.2. The experimental protocols, source and target datasets, evaluated detectors, and rectification settings are described in Sec. 4. All baseline methods are reproduced following their official repositories when available, or otherwise according to the methodological and implementation details reported in the corresponding papers. In particular, all source statistics are estimated before target-domain evaluation, and the same rectification hyperparameters are used without target-specific tuning. Additional implementation details, hyperparameter sensitivity analyses, efficiency evaluations, perturbation studies, and analyses of rectification behavior are provided in the Appendix. These details are intended to facilitate independent reproduction of the reported results.

REFERENCES

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 1–9, 2022.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3291–3300, 2018.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8789–8797, 2018.

Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017.

Peng Cui, Dan Zhang, Zhijie Deng, Yinpeng Dong, and Jun Zhu. Learning sample difficulty from pre-trained models for reliable prediction. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11065–11074, 2019.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. In *International Conference on Learning Representations*, 2020.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*, 2020.

Nicholas Dufour, Andrew Gully, Per Karlsson, Alexey Victor Vorbyov, Thomas Leung, Jeremiah Childs, and Christoph Bregler. Deepfakes detection dataset by google & jigsaw, 2019.

Xinghe Fu, Zhiyuan Yan, Taiping Yao, Shen Chen, and Xi Li. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3040–3048, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

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

Yonghyun Jeong, Doyeon Kim, Seungjai Min, Seongho Joe, Youngjune Gwon, and Jongwon Choi. BiHPF: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 48–57, 2022.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision – ECCV 2022*, volume 13693 of *Lecture Notes in Computer Science*, pp. 709–727. Springer, 2022. doi: 10.1007/978-3-031-19827-4_41.

Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2889–2898, June 2020.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Pavel Korshunov and Sébastien Marcel. DeepFakes: A new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

Neil D. Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in Neural Information Processing Systems*, volume 16, pp. 329–336, 2004.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5001–5010, June 2020a.

Yiheng Li, Zichang Tan, Guoqing Xu, Zhen Lei, Xu Zhou, and Yang Yang. Towards generalizable AI-generated image detection via image-adaptive prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2026.

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

Bo Liu, Qiao Qin, and Qinghui He. Causalclip: Causally-informed feature disentanglement and filtering for generalizable detection of generated images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7069–7077, 2026.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 772–781, June 2021.

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

Andreas Loukas. How close are the eigenvectors of the sample and actual covariance matrices? In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2228–2237, 2017.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 2022.

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

Prashant Pandey, Mrigank Raman, Sumanth Varambally, and Prathosh AP. Generalization on unseen domains via inference-time label-preserving target projections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12924–12933, 2021.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346, 2019.

Wang Qi, Yu-Ping Ruan, Yuan Zuo, and Taihao Li. Parameter-efficient tuning on layer normalization for pre-trained language models. *arXiv preprint arXiv:2211.08682*, 2022.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision – ECCV 2020*, volume 12357 of *Lecture Notes in Computer Science*, pp. 86–103. Springer, 2020. doi: 10.1007/978-3-030-58610-2_6.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Sam T. Roweis. Em algorithms for pca and spca. In *Advances in Neural Information Processing Systems*, volume 10, pp. 626–632, 1998.

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

Anurag Singh, Mahalakshmi Sabanayagam, Krikamol Muandet, and Debarghya Ghoshdastidar. Robust feature inference: A test-time defense strategy using spectral projections. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for GAN-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12105–12114, 2023.

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

Chuangchuang Tan, Ruize Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025.

Linwei Tao, Minjing Dong, and Chang Xu. Feature clipping for uncertainty calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20841–20849, 2025. doi: 10.1609/aaai.v39i19.34297.

Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2387–2395, 2016.

Michael E. Tipping and Christopher M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999. doi: 10.1111/1467-9868.00196.

Gaojian Wang, Feng Lin, Tong Wu, Zhenguang Liu, Zhongjie Ba, and Kui Ren. FSFM: A generalizable face security foundation model via self-supervised facial representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24364–24376, June 2025.

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

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023.

Yuxiang Xu, Rundong He, Zhiyuan Yan, Yicong Dong, Zhongyi Han, Xiaoyan Wang, and Yilong Yin. BPL: Generalizable deepfake detection via bias-only pair-aware learning. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Jiazhen Yan, Ziqiang Li, Fan Wang, Boyu Wang, Ziwen He, and Zhangjie Fu. DGS-Net: Distillation-guided gradient surgery for CLIP fine-tuning in AI-generated image detection. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

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

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Advances in Neural Information Processing Systems*, volume 36, pp. 4534–4565, 2023b.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8984–8994, 2024.

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

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14635–14645, 2025.

Andrii Yermakov, Jan Cech, Jiri Matas, and Mario Fritz. Deepfake detection that generalizes across benchmarks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 773–783, 2026. doi: 10.1109/WACV61042.2026.00082.

Haifeng Zhang, Qinghui He, Xiuli Bi, Weisheng Li, Bo Liu, and Bin Xiao. Towards universal AI-generated image detection by variational information bottleneck network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23828–23837, 2025.

Tianxiang Zhang, Peipeng Yu, Zhihua Xia, Longchen Dai, Xiaoyu Zhou, and Hui Gao. Fine-grained DINO tuning with dual supervision for face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 12780–12788, 2026. doi: 10.1609/aaai.v40i15.38275.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.

Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Shouhong Ding, and Lizhuang Ma. Test-time domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 175–187, June 2024. doi: 10.1109/CVPR52733.2024.00025.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.

Yao Zhu, Yuefeng Chen, Chuanlong Xie, Xiaodan Li, Rong Zhang, Hui Xue, Xiang Tian, Bolun Zheng, and Yaowu Chen. Boosting out-of-distribution detection with typical features. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20758–20769, 2022.

Wanyi Zhuang, Qi Chu, Zhentao Tan, Qiankun Liu, Haojie Yuan, Changtao Miao, Zixiang Luo, and Nenghai Yu. Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European Conference on Computer Vision*, pp. 391–407, 2022.

Yingxuan Zhuang, Jingxiao Yang, Miao Pan, Cheng Tan, Yuxiang Cai, Siwei Tan, Chen Zhi, Xuhong Zhang, Jianwei Yin, and Jintao Chen. Mitigating manifold departure: Uncertainty-aware subspace rectification for trustworthy MLLM decoding. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Appendix

A.1 Additional Details of Preliminary Studies

We provide additional details for preliminary studies in Sec. 1. All detectors are trained on the c23 version of FaceForensics++ (FF++) (Rossler et al., 2019) and evaluated on the unseen DFDCP dataset (Dolhansky et al., 2019). The interpolation coefficient λ is used only for these feasibility studies and is independent of the rectification scale α used by SGSR. No detector parameters are updated during interpolation.

Distance–accuracy analysis. For each detector and each analyzed layer, the reliable source center $\boldsymbol{\mu}$ is computed as the mean representation of correctly predicted FF++-c23 training samples. For each DFDCP test representation $\boldsymbol{z}$, we measure its deviation from the reliable source center using the raw Euclidean distance

$$d(\boldsymbol{z}, \boldsymbol{\mu}) = \|\boldsymbol{z} - \boldsymbol{\mu}\|_2 . \tag{21}$$

Within each layer, let N denote the total number of test samples and let $d_i = d(\boldsymbol{z}_i, \boldsymbol{\mu})$ be the raw Euclidean distance of the i-th test representation to the reliable source center. We sort all test samples in ascending order of d_i, obtaining $d_{(1)} \le d_{(2)} \le \cdots \le d_{(N)}$. The ranked samples are then partitioned into ten approximately equally sized groups according to their rank. Specifically, the k-th decile contains samples with ranks

$$\left\lfloor \frac{(k-1)N}{10} \right\rfloor + 1 \;\text{ to }\; \left\lfloor \frac{kN}{10} \right\rfloor, \qquad k = 1, \ldots, 10. \tag{22}$$

Thus, the first decile contains the 10% of test samples closest to the reliable source center, whereas the tenth decile contains the 10% farthest from it. We compute the detection accuracy independently within each decile,

$$\mathrm{Acc}_k = \frac{1}{|\mathcal{G}_k|} \sum_{i \in \mathcal{G}_k} \mathbb{I}(\hat{y}_i = y_i), \tag{23}$$

where $\mathcal{G}_k$ denotes the set of samples assigned to the k-th decile. This procedure is performed independently for each analyzed layer, yielding the distance–accuracy relation reported in Fig. 1a. This rank-based grouping also facilitates consistent aggregation and comparison of distance–accuracy trends across different layers.

Interpolation toward the reliable source center. For each detector, the reliable source center $\boldsymbol{\mu}$ is computed from correctly predicted FF++-c23 training representations. Given a DFDCP test representation $\boldsymbol{z}$, we perform

$$\boldsymbol{z}' = \boldsymbol{z} + \lambda(\boldsymbol{\mu} - \boldsymbol{z}). \tag{24}$$

We evaluate CLIP-LoRA (Hu et al., 2022), BPL (Xu et al., 2026), Effort (Yan et al., 2025), and GenD (Yermakov et al., 2026). The main-text result in Fig. 1b uses $\lambda = 0.10$.

Table 5: Sensitivity of source-center interpolation to λ. Baseline reports accuracy (%); other entries report ΔAcc.

Detector	Baseline	$\lambda = 0.05$	$\lambda = 0.10$	$\lambda = 0.20$
CLIP-LoRA	76.56	+1.33	+0.39	-5.08
BPL	70.16	+2.73	+5.16	+6.72
Effort	75.90	+1.99	+4.38	-1.20
GenD	76.33	+2.89	+3.83	+3.83
Avg.	74.74	+2.24	+3.44	+1.07

As shown in Table 5, moderate interpolation consistently improves the average performance, with $\lambda = 0.10$ providing the largest average gain. This supports the feasibility of rectifying test representations toward a reliable source region, while also showing that the preferred interpolation strength varies across detectors.

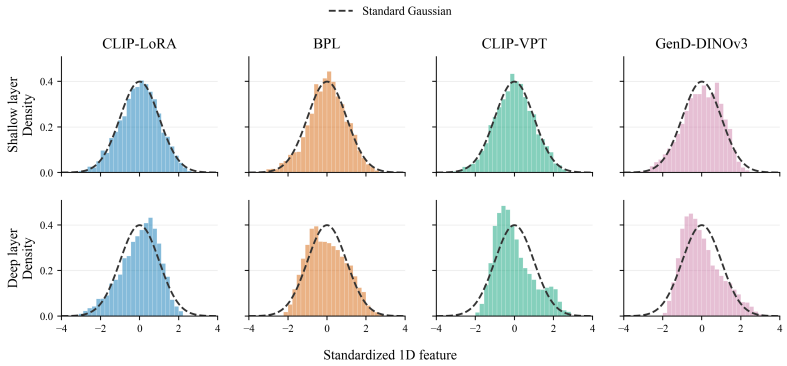


Figure 3: Qualitative visualization of standardized one-dimensional reliable representation distributions.

A.2 Inference Procedure

For completeness, Algorithm 1 summarizes the layer-wise inference procedure of SGSR. All source statistics are estimated offline and remain fixed during inference.

Algorithm 1 Inference with SGSR

Require: Frozen detector f; layer-wise source statistics $\{\Theta_\ell\}$; rectification parameter τ_ℓ^2

1: Initialize the detector with input image $\boldsymbol{x}$
2: **for** each intermediate layer ℓ in forward order **do**
3: Compute the intermediate state at layer ℓ
4: Extract the global representation $\boldsymbol{z}_\ell$
5: $\boldsymbol{c} \leftarrow \mathbf{U}_{K,\ell}^\top(\boldsymbol{z}_\ell - \boldsymbol{\mu}_\ell)$
6: $\boldsymbol{r} \leftarrow (\boldsymbol{z}_\ell - \boldsymbol{\mu}_\ell) - \mathbf{U}_{K,\ell}\boldsymbol{c}$
7: $\hat{\boldsymbol{u}}_\ell \leftarrow \boldsymbol{\mu}_\ell + \mathbf{U}_{K,\ell}\,\mathrm{diag}\left(\frac{\lambda_{\ell,i}}{\lambda_{\ell,i}+\tau_\ell^2}\right)\boldsymbol{c} + \frac{\sigma_{\perp,\ell}^2}{\sigma_{\perp,\ell}^2+\tau_\ell^2}\boldsymbol{r}$
8: Reinsert $\hat{\boldsymbol{u}}_\ell$ into the intermediate state
9: **end for**
10: **return** detector prediction

A.3 Qualitative Visualization of the Gaussian Prior

We provide a qualitative visualization of the Gaussian prior used in Sec. 3.2. We stress that this analysis is not meant to prove that reliable representations strictly follow a Gaussian distribution. Rather, it serves to examine whether Gaussian modeling provides a reasonable approximation to the observed representation behavior.

Using correctly predicted source-domain representations from the FF++-c23 training set, we visualize the empirical standardized one-dimensional representation distributions for CLIP-LoRA (Hu et al., 2022), BPL (Xu et al., 2026), CLIP-VPT (Jia et al., 2022), and GenD-DINOv3 (Yermakov et al., 2026). For each detector and analyzed layer, PCA is fitted on the corresponding reliable source representations, and each representation is projected onto the first principal component. Figure 3 shows examples from a shallow layer (layer 6) and a deep layer (layer 16), together with a standard Gaussian reference.

The empirical distributions are generally unimodal and approximately bell-shaped, especially in the shallower layer, while deeper layers show some skewness and heavier tails. These observations suggest that Gaussian modeling is not exact, but remains a reasonable low-complexity approximation for reliable representation rectification.

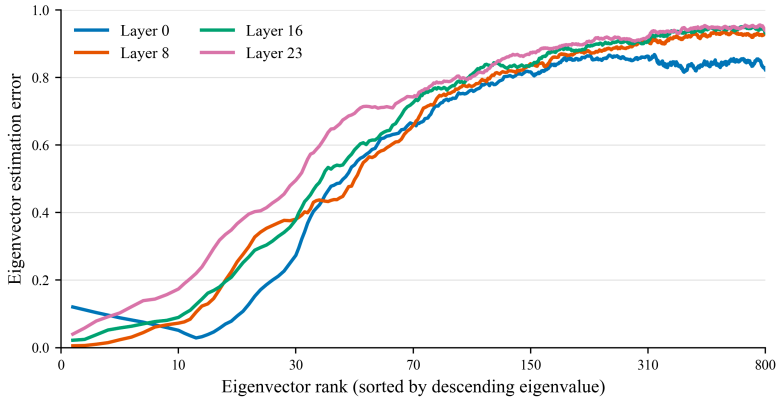


Figure 4: **Stability of covariance eigendirection estimation.** Covariance matrices are independently estimated from ten subsets of the FF++ training set using CLIP-LoRA `[CLS]` representations. Eigendirections are ranked by descending eigenvalue, and estimation error is measured relative to the eigendirections obtained from the full training set. Tail eigendirections exhibit substantially larger estimation errors across different Transformer layers.

A.4 Stability of Covariance Eigenspace Estimation

To empirically examine the stability of covariance estimation under finite source samples, we conduct a split-based analysis using CLIP-LoRA on FaceForensics++ (FF++). We extract the `[CLS]` representations of correctly predicted source samples and estimate a reference covariance matrix using the full FF++ training set. We then randomly partition the training set into 10 disjoint subsets and independently estimate the covariance matrix from each subset.

For each layer l, the eigendirections are ordered by their corresponding eigenvalues in descending order. Let $\mathbf{u}_{l,i}^{\text{full}}$ denote the i-th eigendirection estimated from the full training set and $\mathbf{u}_{l,i}^{(k)}$ the corresponding eigendirection estimated from the k-th subset. We measure the rank-wise eigendirection estimation error as

$$e_{l,i} = \frac{1}{10} \sum_{k=1}^{10} \left(1 - \left| {\mathbf{u}_{l,i}^{(k)}}^{\top} \mathbf{u}_{l,i}^{\text{full}} \right| \right). \tag{25}$$

Since all eigendirections are unit-normalized, the absolute inner product corresponds to the absolute cosine similarity, removing the sign ambiguity of eigenvectors. Thus, a smaller $e_{l,i}$ indicates more stable eigendirection estimation across different source subsets.

As shown in Fig. 4, the estimation error remains relatively low for the leading eigendirections but increases substantially as the eigenvector rank grows. This trend is consistently observed across different Transformer layers. In particular, the low-rank principal directions, which correspond to larger eigenvalues, can be estimated substantially more consistently, whereas the tail eigendirections exhibit much larger variations under different source subsets. These results empirically support our low-rank covariance modeling: SGSR preserves the dominant principal directions with more stable estimates, while summarizing the less stable tail directions using the shared residual variance $\sigma_{\perp}^2$.

A.5 Efficiency Analysis

We first evaluate the overall storage and inference overhead introduced by SGSR on two representative architectures, CLIP-LoRA and ResNet-50. All runtime measurements are conducted on a single NVIDIA V100 GPU with a batch size of 32.

Table 8: Sensitivity to the rectification scale α on Effort and ResNet-50. The baseline row reports the original AUC (%), while all other rows report ΔAUC relative to the corresponding baseline. The PCA explained-variance ratio is fixed to $\eta = 0.9$.

	Effort						ResNet-50					
α	DFD	CDF-v2	DFDCP	DFDC	WDF	Avg.	DFD	CDF-v2	DFDCP	DFDC	WDF	Avg.
Baseline	93.710	85.622	85.272	81.315	86.123	86.408	78.160	61.285	63.848	65.413	74.321	68.605
0.05	+0.289	+0.367	+0.463	+0.835	+1.137	+0.618	+0.398	+0.157	+0.122	+0.313	-0.385	+0.121
0.10	+0.382	+0.385	+0.453	+1.022	+1.646	+0.778	+0.663	+0.288	+0.143	+0.538	-0.725	+0.181
0.20	+0.368	+0.460	+0.298	+1.127	+2.224	**+0.895**	+0.998	+0.517	+0.058	+0.843	-1.333	**+0.217**
0.40	+0.389	+0.244	+0.124	+0.978	+2.364	+0.820	+1.324	+0.872	-0.250	+1.171	-2.360	+0.151
0.80	+0.353	+0.090	-0.094	+0.852	+2.480	+0.736	+1.481	+1.335	-0.866	+1.407	-3.978	-0.124

Table 6: Overall efficiency analysis of SGSR. "Params" reports the detector parameter count and the additional source-estimated statistics stored by SGSR.

Detector	Params (Base / SGSR)	Storage Overhead	Latency (Base / SGSR)	Time Overhead
CLIP-LoRA	306.33M / 3.32M	+1.08%	514.72 / 518.02 ms	+0.64%
ResNet-50	23.51M / 28.77K	+0.12%	37.25 / 38.59 ms	+3.60%

As shown in Table 6, SGSR introduces only lightweight storage and computational overhead. The stored statistics correspond to approximately 1.08% and 0.12% of the parameter counts of CLIP-LoRA and ResNet-50, respectively, while inference latency increases by only 0.64% and 3.60%.

We further compare the inference efficiency of the proposed low-rank rectification with its full-covariance counterpart in Eq. 9. Both formulations are applied at all intermediate layers under the same evaluation setting.

Table 7: Inference efficiency of low-rank and full-covariance rectification. Overhead is measured relative to the original detector.

Detector	Baseline	Low-Rank	Full-Cov	Low-Rank Overhead	Full-Cov Overhead
CLIP-LoRA	514.72 ms	518.02 ms	545.03 ms	+0.64%	+5.89%
ResNet-50	37.25 ms	38.59 ms	41.44 ms	+3.60%	+11.22%

As shown in Table 7, the low-rank formulation introduces less inference overhead than full-covariance rectification. On CLIP-LoRA, the overhead is reduced from 5.89% to 0.64%, while on ResNet-50 it decreases from 11.22% to 3.60%. Equivalently, full-covariance rectification is 5.21% and 7.39% slower than its low-rank counterpart on the two detectors, respectively. These results empirically support the computational advantage of the proposed low-rank formulation, which avoids dense covariance operations and reduces the per-layer complexity from $\mathcal{O}(d^2)$ to $\mathcal{O}(dK)$ for $K \ll d$.

A.6 Hyperparameter Sensitivity.

Tables 8 and 9 analyze the sensitivity of SGSR to the rectification scale α and the PCA explained-variance ratio η on two representative architectures, Effort (Yan et al., 2025) and ResNet-50 (He et al., 2016). The rectification scale α is selected using only source-domain validation data, while $\eta = 0.9$ is fixed as a predefined explained-variance criterion. Both are fixed before target-domain evaluation and used consistently across all detectors and unseen target domains. The results in this section are reported only to examine sensitivity to these choices and are not used for target-specific hyperparameter selection. When varying one hyperparameter, the other is fixed to its default value.

For α, moderate rectification scales generally yield stable positive improvements on both architectures, whereas overly large values can degrade performance on some target datasets. Effort remains relatively stable over $\alpha \in [0.1, 0.8]$, while ResNet-50 becomes more sensitive as the rectification

Table 9: Sensitivity to the PCA explained-variance ratio η on Effort and ResNet-50. The baseline row reports the original AUC (%), while all other rows report ΔAUC relative to the corresponding baseline. The rectification scale is fixed to $\alpha = 0.2$.

	Effort						ResNet-50					
η	DFD	CDF-v2	DFDCP	DFDC	WDF	Avg.	DFD	CDF-v2	DFDCP	DFDC	WDF	Avg.
Baseline	93.710	85.622	85.272	81.315	86.123	86.408	78.160	61.285	63.848	65.413	74.321	68.605
0.70	+0.432	+0.724	+0.531	+1.141	+2.016	+0.969	+0.748	+1.019	-0.502	+0.750	-2.336	-0.064
0.80	+0.351	+0.820	+0.541	+1.131	+2.037	**+0.976**	+0.801	+0.821	-0.417	+0.791	-1.666	+0.066
0.90	+0.368	+0.460	+0.298	+1.127	+2.224	+0.895	+0.998	+0.517	+0.058	+0.843	-1.333	+0.217
0.95	+0.341	+0.588	+0.079	+1.025	+1.906	+0.788	+1.082	+0.630	+0.216	+0.701	-0.637	**+0.398**
0.99	+0.394	+0.205	-0.191	+0.919	+1.674	+0.600	+1.098	+0.538	+0.713	+0.320	-0.764	+0.381

Table 10: Comparison with generic inference-time representation intervention methods on four representative detectors. Each entry reports ΔAUC (%) relative to the corresponding original detector. "Avg." denotes the average over the five unseen deepfake datasets.

Detector	Method	DFD	CDF-v2	DFDCP	DFDC	WDF	Avg.
CLIP-LoRA	Feature Clipping (Tao et al., 2025)	-0.092	+0.064	+0.164	-0.177	-0.786	-0.165
	ReAct (Sun et al., 2021)	+0.075	+0.186	+0.020	-0.025	+0.281	+0.107
	RFI (Singh et al., 2024)	+1.542	+1.679	+0.733	+1.774	+2.638	+1.673
ResNet-50	Feature Clipping (Tao et al., 2025)	-2.246	-2.115	-2.676	+0.263	+0.757	-1.203
	ReAct (Sun et al., 2021)	-3.115	-2.695	-3.684	+0.172	+0.913	-1.682
	RFI (Singh et al., 2024)	+0.192	+0.384	+0.380	-0.038	-0.091	+0.165
GenD-DINOv3	Feature Clipping (Tao et al., 2025)	-0.022	-0.095	+0.061	-0.067	+0.224	+0.020
	ReAct (Sun et al., 2021)	+0.034	-0.071	+0.007	+0.011	+0.103	+0.017
	RFI (Singh et al., 2024)	-0.004	-0.378	-0.213	-0.202	-0.113	-0.182
Effort	Feature Clipping (Tao et al., 2025)	-0.019	+0.116	+0.247	+0.104	-0.078	+0.074
	ReAct (Sun et al., 2021)	-0.190	+0.194	-0.197	-0.077	-0.072	-0.068
	RFI (Singh et al., 2024)	-1.412	-10.938	-7.143	-1.952	-0.408	-4.371
Macro Avg. – Feature Clipping		-0.595	-0.508	-0.551	+0.031	+0.029	-0.319
Macro Avg. – ReAct		-0.799	-0.597	-0.964	+0.020	+0.306	-0.407
Macro Avg. – RFI		+0.080	-2.313	-1.561	-0.105	+0.507	-0.679

scale increases. For η, the two architectures exhibit different sensitivity patterns. Effort favors a relatively compact principal subspace, whereas ResNet-50 benefits from retaining more explained variance. Overall, SGSR remains effective over a reasonable range of hyperparameter values around the default configuration. These results indicate that the observed improvements do not depend on fine-grained target-specific tuning.

A.7 Comparison with Generic Inference-Time Representation Interventions

To further examine whether generic inference-time representation modification is sufficient to improve cross-domain forgery detection, we evaluate three representative representation intervention methods on four detectors: Feature Clipping (Tao et al., 2025), ReAct (Sun et al., 2021), and Robust Feature Inference (RFI) (Singh et al., 2024). These methods were originally developed for uncertainty calibration, out-of-distribution detection, and adversarial robustness, respectively, rather than cross-domain AI-generated image detection. We therefore treat them as generic inference-time representation intervention baselines rather than task-specific competing methods.

As shown in Table 10, the three generic representation intervention methods exhibit markedly different behavior across detectors. Feature Clipping and ReAct provide small improvements in some settings but produce negative macro-average gains of -0.319 and -0.407 AUC points, respectively. RFI achieves a substantial improvement on CLIP-LoRA but degrades GenD-DINOv3 and, in particular, Effort, resulting in an overall average change of -0.679 points.

These results indicate that inference-time representation modification alone does not necessarily improve cross-domain forgery generalization. Existing intervention strategies are designed around different objectives—confidence calibration, OOD activation suppression, or adversarial robustness—and therefore do not explicitly model the reliable representation geometry relevant to forgery detection. In contrast, SGSR uses source-estimated reliable representation statistics to perform direction-dependent representation rectification, providing an alternative inference-time intervention strategy for cross-domain AIGI detection.

A.8 Robustness under Image Perturbations

We further investigate whether SGSR can enhance detector robustness under common image perturbations. Following prior perturbation-based robustness evaluations (Jiang et al., 2020), we consider four representative perturbations, including Gaussian blur, JPEG compression, Gaussian noise, and block-wise distortion, each with five severity levels. We evaluate four detectors with different architectures, including ResNet-50 (He et al., 2016), CLIP-LoRA (Hu et al., 2022), UIA-ViT (Zhuang et al., 2022), and Effort (Yan et al., 2025). All detector parameters and source-estimated statistics remain fixed throughout evaluation.

For clarity, Table 11 reports the average ΔAUC over the five severity levels for each perturbation, while Table 12 summarizes the average performance at each severity level. The full set of individual results is omitted for presentation clarity.

Table 11: Robustness under different image perturbations. Each entry reports the mean ΔAUC over five severity levels.

Detector	Blur	JPEG	Noise	Block	Avg.
ResNet-50	+0.692	+1.104	+1.912	+0.653	+1.090
CLIP-LoRA	+0.143	+0.288	+0.608	+1.267	+0.576
UIA-ViT	+0.086	-0.013	+0.283	+0.076	+0.108
Effort	+0.133	+0.503	+1.518	+4.499	+1.663
Avg.	+0.263	+0.470	+1.080	+1.624	+0.859

Table 12: Average ΔAUC across all detectors and perturbation types at different severity levels.

Severity	1	2	3	4	5
Avg. ΔAUC	+0.355	+0.625	+0.886	+1.190	+1.241

As shown in Tables 11 and 12, SGSR provides overall positive improvements across diverse perturbation types and detector architectures. The average gain reaches 0.859 AUC points, with more pronounced improvements under Gaussian noise and block artifacts. Moreover, the average benefit increases from 0.355 AUC points at severity 1 to 1.241 points at severity 5. Although this trend is not strictly monotonic for every individual setting, the aggregate results suggest that SGSR tends to provide greater benefits under stronger image perturbations.

A.9 Additional Cross-Generator Evaluation on AIGCDetectBenchmark

Evaluation tasks and datasets. To further evaluate the generalizability of SGSR beyond the UniversalFakeDetect (UFD) benchmark, we conduct an additional cross-generator evaluation on AIGCDetectBenchmark (Zhong et al., 2023). We use the same detectors and follow the same training and rectification settings as in Sec. 4.2. To avoid duplicate evaluation with UFD, we exclude BigGAN, CycleGAN, GauGAN, ProGAN, StarGAN, StyleGAN, and GLIDE, which are already covered by the main synthetic-image evaluation. The remaining ten subsets include ADM, DALL-E 2, Midjourney, VQDM, SDXL, Stable Diffusion 1.4, Stable Diffusion 1.5, StyleGAN2, WhichFaceIsReal, and Wukong. As in the main experiments, the source statistics required by SGSR are

Table 13: Cross-generator synthetic image detection AUC (%) on ten non-overlapping AIGCDetect-Benchmark subsets. Subsets overlapping with the UFD evaluation are excluded. Each entry reports baseline AUC (ΔAUC), with positive improvements shown in bold.

Method	ADM	DALL-E 2	Midjourney	VQDM	SDXL
CLIP-FFT (Radford et al., 2021)	82.92 (**+12.72**)	58.42 (**+34.99**)	71.46 (**+14.55**)	95.34 (**+3.13**)	84.05 (**+11.61**)
CLIP-LN (Qi et al., 2022)	89.72 (**+7.55**)	86.48 (**+12.47**)	68.79 (**+10.08**)	96.79 (**+1.95**)	79.34 (**+15.84**)
CLIP-LoRA (Hu et al., 2022)	82.85 (**+8.10**)	94.20 (**+5.27**)	40.76 (**+15.68**)	93.01 (**+4.75**)	69.74 (**+21.80**)
CLIP-VPT (Jia et al., 2022)	91.34 (**+2.65**)	86.14 (**+11.03**)	78.44 (**+0.44**)	97.86 (-0.52)	92.48 (**+3.86**)
CNNSpot (Wang et al., 2020)	82.91 (**+0.25**)	79.31 (**+1.65**)	60.92 (**+2.47**)	72.26 (**+0.46**)	68.81 (-1.84)
CLIP-BitFit (Ben Zaken et al., 2022)	86.20 (**+8.37**)	86.74 (**+12.38**)	60.04 (**+4.29**)	94.47 (**+2.94**)	81.66 (**+12.22**)
UniFD (Ojha et al., 2023)	91.67 (**+1.65**)	76.11 (**+6.04**)	66.03 (**+12.54**)	96.90 (-0.09)	83.95 (**+7.56**)
NPR (Tan et al., 2024)	90.11 (**+1.59**)	94.32 (**+1.86**)	74.87 (**+2.75**)	89.65 (**+1.76**)	95.44 (**+1.13**)
C2P-CLIP (Tan et al., 2025)	91.50 (**+4.47**)	97.17 (**+2.48**)	76.76 (**+2.31**)	97.65 (-0.09)	89.06 (**+6.49**)
VIB-Net (Zhang et al., 2025)	84.98 (**+7.56**)	72.08 (**+9.20**)	66.12 (**+13.32**)	93.71 (**+2.43**)	72.11 (**+18.24**)
Effort (Yan et al., 2025)	86.76 (**+8.44**)	69.28 (**+29.29**)	62.92 (**+13.01**)	95.22 (**+1.60**)	75.22 (**+16.12**)
GenD-DINOv3 (Yermakov et al., 2026)	79.40 (**+5.17**)	80.11 (**+0.12**)	58.62 (-8.90)	90.08 (**+3.26**)	74.53 (-1.65)
CausalCLIP (Liu et al., 2026)	87.64 (**+4.80**)	69.54 (**+11.56**)	70.89 (**+9.44**)	95.73 (**+0.86**)	81.53 (**+10.37**)
Avg.	86.77 (**+5.64**)	80.76 (**+10.64**)	65.89 (**+7.08**)	92.97 (**+1.73**)	80.61 (**+9.37**)

Method	SD 1.4	SD 1.5	StyleGAN2	WhichFace	Wukong	Avg.
CLIP-FFT (Radford et al., 2021)	89.24 (**+9.31**)	89.19 (**+9.08**)	97.13 (**+2.68**)	97.47 (**+1.69**)	90.67 (**+7.75**)	85.59 (**+10.75**)
CLIP-LN (Qi et al., 2022)	96.48 (**+2.54**)	96.45 (**+2.34**)	99.88 (**+0.09**)	98.39 (**+1.05**)	96.57 (**+2.15**)	90.89 (**+5.61**)
CLIP-LoRA (Hu et al., 2022)	94.59 (**+4.86**)	93.90 (**+5.29**)	98.71 (**+1.09**)	91.62 (**+7.74**)	93.61 (**+5.38**)	85.30 (**+8.00**)
CLIP-VPT (Jia et al., 2022)	98.57 (-0.70)	98.21 (-0.54)	99.87 (**+0.09**)	99.82 (-1.05)	97.92 (-0.98)	94.06 (**+1.43**)
CLIP-BitFit (Ben Zaken et al., 2022)	96.42 (**+2.63**)	95.91 (**+2.83**)	99.73 (**+0.24**)	98.22 (**+0.95**)	96.43 (**+2.40**)	89.58 (**+4.93**)
CNNSpot (Wang et al., 2020)	84.30 (-0.39)	84.01 (-0.47)	95.75 (**+0.60**)	92.56 (**+0.89**)	77.00 (-0.13)	79.78 (**+0.35**)
UniFD (Ojha et al., 2023)	90.65 (**+2.86**)	90.52 (**+2.61**)	98.88 (**+0.88**)	98.06 (-0.13)	93.61 (**+1.26**)	88.64 (**+3.52**)
NPR (Tan et al., 2024)	91.42 (**+1.07**)	91.05 (**+1.13**)	99.91 (**+0.00**)	63.88 (-0.03)	84.98 (**+1.38**)	87.56 (**+1.27**)
C2P-CLIP (Tan et al., 2025)	98.64 (**+0.78**)	98.57 (**+0.76**)	99.28 (-0.50)	99.26 (-1.36)	98.00 (**+0.97**)	94.59 (**+1.63**)
VIB-Net (Zhang et al., 2025)	83.46 (**+10.14**)	83.38 (**+9.83**)	93.38 (**+4.68**)	97.38 (-0.22)	88.74 (**+5.93**)	83.53 (**+8.11**)
Effort (Yan et al., 2025)	94.25 (**+3.93**)	94.42 (**+3.51**)	98.73 (**+0.43**)	95.51 (**+1.96**)	93.63 (**+4.12**)	86.59 (**+8.24**)
GenD-DINOv3 (Yermakov et al., 2026)	85.90 (**+3.55**)	85.94 (**+2.89**)	97.27 (**+1.65**)	58.04 (**+33.29**)	85.83 (**+3.08**)	79.57 (**+4.25**)
CausalCLIP (Liu et al., 2026)	90.27 (**+3.80**)	90.10 (**+3.62**)	98.86 (**+0.83**)	98.26 (-0.47)	93.05 (**+2.06**)	87.59 (**+4.68**)
Avg.	91.86 (**+3.41**)	91.67 (**+3.30**)	98.26 (**+0.98**)	91.42 (**+3.41**)	91.54 (**+2.72**)	87.18 (**+4.83**)

estimated before target-domain evaluation, and no target-domain labels or statistics are used during rectification.

Evaluation results. As shown in Table 13, SGSR improves the overall average AUC by 4.83 percentage points across 13 detectors and ten additional generator subsets, with positive average gains for all evaluated detectors.

A.10 Analysis of Rectification Behavior

We further analyze the rectification behavior of SGSR using four representative detectors, including ResNet-50 (He et al., 2016), Effort (Yan et al., 2025), GenD-CLIP-L (Yermakov et al., 2026), and BPL (Xu et al., 2026), following the main cross-dataset setting. Specifically, we first report a post-hoc observation on their representation variance, and then analyze the relationship between group-level representation statistics and the magnitude of the representation- and prediction-level changes induced by SGSR. Target labels are used only for post-hoc analysis and are never involved in source-statistics estimation or inference-time rectification.

Post-hoc Observation on Target Representation Variance. Let $g \in \{C, W\}$ denote the correctly predicted (C) and wrongly predicted (W) target samples, respectively. We characterize the representation variance of each group by

$$\sigma_g^2 = \frac{1}{d} \operatorname{Tr}\left[\operatorname{Cov}(\boldsymbol{z} \mid g)\right]. \tag{26}$$

The reported target representation variances are class-balanced over real and fake samples.

Table 14: Representation variance of correctly (C) and wrongly (W) predicted target samples, averaged over the five unseen Deepfake datasets.

Detector	σ_C^2	σ_W^2	σ_W^2/σ_C^2
ResNet-50	3.82×10^{-4}	4.02×10^{-4}	1.051
Effort	0.01585	0.01618	1.021
GenD-CLIP-L	0.02257	0.02293	1.016
BPL	0.03886	0.05927	1.525

As shown in Table 14, wrongly predicted target representations consistently exhibit larger representation variance than correctly predicted representations across all four representative detectors:

$$\sigma_W^2 > \sigma_C^2. \tag{27}$$

We emphasize that this is a post-hoc empirical observation rather than a causal claim: larger representation variance is associated with wrong predictions, but is not assumed to cause prediction errors.

Representation-Level Rectification Behavior. We next analyze how representation variance affects the rectification magnitude of SGSR. The rectification in Eq. 15 can be compactly written as

$$\hat{\boldsymbol{u}} = \boldsymbol{\mu} + \mathbf{A}(\boldsymbol{z} - \boldsymbol{\mu}), \tag{28}$$

where

$$\mathbf{A} = \mathbf{U}_K \operatorname{diag}\left(\frac{\lambda_i}{\lambda_i + \tau^2}\right) \mathbf{U}_K^\top + \frac{\sigma_\perp^2}{\sigma_\perp^2 + \tau^2}\left(\mathbf{I} - \mathbf{U}_K \mathbf{U}_K^\top\right). \tag{29}$$

Let

$$\mathbf{B} = \mathbf{I} - \mathbf{A}. \tag{30}$$

The representation modification introduced by SGSR is therefore

$$\Delta\boldsymbol{z} = \hat{\boldsymbol{u}} - \boldsymbol{z} = -\mathbf{B}(\boldsymbol{z} - \boldsymbol{\mu}). \tag{31}$$

To analyze the relationship between representation variance and rectification magnitude, we consider a group-dependent extension of the Gaussian observation model in Sec. 3.2. This group-dependent model is introduced only as a simplified analytical abstraction for characterizing the observed rectification behavior, rather than as an exact generative model of the Correct and Wrong groups. For each group $g \in \{C, W\}$, we write

$$\boldsymbol{z} = \boldsymbol{u} + \boldsymbol{\epsilon}_g, \qquad \boldsymbol{\epsilon}_g \sim \mathcal{N}\left(\mathbf{0}, \rho_g^2 \mathbf{I}\right), \tag{32}$$

where $\boldsymbol{\epsilon}_g$ is independent of the latent reliable representation, which follows

$$\boldsymbol{u} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}_r\right). \tag{33}$$

Here, ρ_g^2 denotes the perturbation variance of group g and is introduced only for analyzing group-dependent target behavior. It is distinct from the rectification parameter τ^2 used by SGSR.

Under this model,

$$\operatorname{Cov}(\boldsymbol{z} \mid g) = \boldsymbol{\Sigma}_r + \rho_g^2 \mathbf{I}. \tag{34}$$

In particular, this analysis assumes that the two groups share the same underlying reliable covariance structure and differ only in the magnitude of isotropic perturbation. Therefore, the empirically measured group representation variance satisfies

$$\begin{aligned} \sigma_g^2 &= \frac{1}{d} \operatorname{Tr}\left[\operatorname{Cov}(\boldsymbol{z} \mid g)\right] \\ &= \frac{1}{d} \operatorname{Tr}\left(\boldsymbol{\Sigma}_r\right) + \rho_g^2. \end{aligned} \tag{35}$$

Since the reliable representation distribution is shared across the two groups, the empirical observation $\sigma_W^2 > \sigma_C^2$ implies

$$\rho_W^2 > \rho_C^2. \tag{36}$$

We define the representation rectification energy of group g as

$$\mathcal{C}_g = \mathbb{E}_{\boldsymbol{z}|g}\left[\|\hat{\boldsymbol{u}} - \boldsymbol{z}\|_2^2\right]. \tag{37}$$

Using Eq. 31, we obtain

$$\mathcal{C}_g = \mathbb{E}_g\left[(\boldsymbol{z}-\boldsymbol{\mu})^\top \mathbf{B}^\top \mathbf{B}(\boldsymbol{z}-\boldsymbol{\mu})\right] = \mathrm{Tr}\left[\mathbf{B}^\top \mathbf{B}\,\mathrm{Cov}(\boldsymbol{z} \mid g)\right]. \tag{38}$$

Since $\mathbf{A}$ and $\mathbf{B}$ are symmetric, $\mathbf{B}^\top\mathbf{B} = \mathbf{B}^2$. Substituting Eq. 34 gives

$$\mathcal{C}_g = \mathrm{Tr}\left[\mathbf{B}^2\left(\boldsymbol{\Sigma}_r + \rho_g^2\mathbf{I}\right)\right] = \mathrm{Tr}\left(\mathbf{B}^2\boldsymbol{\Sigma}_r\right) + \rho_g^2\,\mathrm{Tr}\left(\mathbf{B}^2\right). \tag{39}$$

Hence,

$$\frac{\partial \mathcal{C}_g}{\partial \rho_g^2} = \mathrm{Tr}\left(\mathbf{B}^2\right) > 0. \tag{40}$$

The inequality is strict for any non-trivial rectification operator $\mathbf{B} \neq \mathbf{0}$.

Furthermore,

$$\mathcal{C}_W - \mathcal{C}_C = \left(\rho_W^2 - \rho_C^2\right)\mathrm{Tr}\left(\mathbf{B}^2\right) = \left(\sigma_W^2 - \sigma_C^2\right)\mathrm{Tr}\left(\mathbf{B}^2\right). \tag{41}$$

Therefore,

$$\sigma_W^2 > \sigma_C^2 \quad \Longrightarrow \quad \mathcal{C}_W > \mathcal{C}_C. \tag{42}$$

This result shows that, under the observation model of SGSR, representations with larger perturbation variance naturally receive stronger representation-level rectification.

The empirical metric below is not identical to the squared rectification energy derived above, but is used to examine whether the same qualitative ordering holds in practice. We further examine this behavior empirically using the mean representation rectification magnitude

$$M_g = \mathbb{E}_{\boldsymbol{z}|g}\left[\|\hat{\boldsymbol{u}} - \boldsymbol{z}\|_2\right]. \tag{43}$$

Table 15: Representation rectification magnitude for correctly (C) and wrongly (W) predicted target samples, averaged over the five unseen Deepfake datasets.

Detector	M_C	M_W	M_W/M_C
ResNet-50	0.00464	0.00479	1.032
Effort	0.06918	0.07206	1.042
GenD-CLIP-L	0.08210	0.08706	1.060
BPL	0.08663	0.09368	1.081

As shown in Table 15, the Wrong group consistently receives larger representation modifications than the Correct group across all four detectors, showing the same ordering predicted by the theoretical analysis.

Prediction-Level Rectification Behavior. Next, we analyze how representation rectification affects the detector prediction. We first consider the case where rectification is applied immediately before a linear classification head:

$$s(\boldsymbol{z}) = \mathbf{w}^\top \boldsymbol{z} + b, \tag{44}$$

where $s(\boldsymbol{z})$ denotes the classification logit.

The logit change introduced by SGSR is

$$\Delta s = s(\hat{\boldsymbol{u}}) - s(\boldsymbol{z}) = \mathbf{w}^\top(\hat{\boldsymbol{u}} - \boldsymbol{z}) = -\mathbf{w}^\top\mathbf{B}(\boldsymbol{z} - \boldsymbol{\mu}). \tag{45}$$

We define the prediction correction energy of group g as

$$\mathcal{P}_g = \mathbb{E}_{\boldsymbol{z}|g}\left[(\Delta s)^2\right]. \tag{46}$$

Under the group-dependent Gaussian perturbation model in Eq. 34, we obtain

$$\begin{aligned}\mathcal{P}_g &= \mathbf{w}^\top \mathbf{B}\left(\mathbf{\Sigma}_r + \rho_g^2 \mathbf{I}\right)\mathbf{B}\mathbf{w} \\ &= \mathbf{w}^\top \mathbf{B}\mathbf{\Sigma}_r \mathbf{B}\mathbf{w} + \rho_g^2 \left\|\mathbf{B}\mathbf{w}\right\|_2^2 .\end{aligned} \tag{47}$$

Therefore,

$$\frac{\partial \mathcal{P}_g}{\partial \rho_g^2} = \left\|\mathbf{B}\mathbf{w}\right\|_2^2 \geq 0. \tag{48}$$

When the rectified representation directions are not completely orthogonal to the classifier direction, i.e.,

$$\mathbf{B}\mathbf{w} \neq \mathbf{0}, \tag{49}$$

the inequality in Eq. 48 is strict. Since Eq. 35 establishes the same ordering between σ_g^2 and ρ_g^2, we obtain

$$\sigma_W^2 > \sigma_C^2 \quad \Longrightarrow \quad \mathcal{P}_W > \mathcal{P}_C. \tag{50}$$

Therefore, when rectification is applied directly before a linear classifier, representations with larger perturbation variance are expected to induce larger prediction-level changes.

For intermediate Transformer or CNN layers, the downstream mapping h_ℓ is generally nonlinear. Let $\boldsymbol{z}_\ell$ and $\hat{\boldsymbol{u}}_\ell$ denote the representations before and after rectification at layer ℓ. A first-order Taylor approximation gives

$$h_\ell\left(\hat{\boldsymbol{u}}_\ell\right) - h_\ell\left(\boldsymbol{z}_\ell\right) \approx \mathbf{J}_\ell\left(\hat{\boldsymbol{u}}_\ell - \boldsymbol{z}_\ell\right), \tag{51}$$

where $\mathbf{J}_\ell$ denotes the local Jacobian of the downstream mapping evaluated around $\boldsymbol{z}_\ell$.

Under a locally fixed Jacobian approximation, we define the local prediction-response energy as

$$\mathcal{P}_{g,\ell} = \mathbb{E}_{\boldsymbol{z}_\ell|g}\left[\left\|h_\ell\left(\hat{\boldsymbol{u}}_\ell\right) - h_\ell\left(\boldsymbol{z}_\ell\right)\right\|_2^2\right]. \tag{52}$$

Using the first-order approximation in Eq. 51, we obtain

$$\begin{aligned}\mathcal{P}_{g,\ell} &\approx \mathrm{Tr}\left[\mathbf{J}_\ell \mathbf{B}_\ell\left(\mathbf{\Sigma}_{r,\ell} + \rho_g^2 \mathbf{I}\right)\mathbf{B}_\ell \mathbf{J}_\ell^\top\right] \\ &= \mathrm{Tr}\left[\mathbf{J}_\ell \mathbf{B}_\ell \mathbf{\Sigma}_{r,\ell} \mathbf{B}_\ell \mathbf{J}_\ell^\top\right] \\ &\quad + \rho_g^2 \left\|\mathbf{J}_\ell \mathbf{B}_\ell\right\|_F^2 .\end{aligned} \tag{53}$$

Thus,

$$\frac{\partial \mathcal{P}_{g,\ell}}{\partial \rho_g^2} \approx \left\|\mathbf{J}_\ell \mathbf{B}_\ell\right\|_F^2 \geq 0. \tag{54}$$

Therefore, under the local linearization of the downstream network, larger perturbation variance also leads to a larger prediction-level response.

To empirically examine this behavior, we report the mean absolute prediction change

$$P_g = \mathbb{E}_{\boldsymbol{z}|g}\left[\left|p_{\mathrm{SGSR}} - p_{\mathrm{Base}}\right|\right]. \tag{55}$$

where p_{Base} and p_{SGSR} denote the prediction probabilities before and after rectification, respectively.

Table 16: Mean absolute prediction change for correctly (C) and wrongly (W) predicted target samples, averaged over the five unseen Deepfake datasets. Values are reported in percentage points.

Detector	P_C	P_W	P_W/P_C
ResNet-50	3.296	4.215	1.279
Effort	5.145	11.434	2.222
GenD-CLIP-L	7.417	19.551	2.636
BPL	5.047	12.855	2.547

As shown in Table 16, wrongly predicted samples consistently undergo larger prediction changes after rectification. The difference is particularly pronounced for Effort, GenD-CLIP-L, and BPL, where the prediction change of the Wrong group is approximately 2.2–2.6$\times$ that of the Correct group. These results are consistent with the theoretical prediction that higher-variance representations induce stronger prediction-level responses.

Prediction Recovery and Degradation. The above analysis characterizes the magnitude of the prediction response but does not determine whether the resulting change improves or degrades the prediction. We therefore further examine the actual prediction transitions induced by SGSR.

We define the recovery rate (RR) as the proportion of originally wrongly predicted samples that become correctly predicted after rectification:

$$\mathrm{RR} = \frac{N_{W\to C}}{N_W}, \tag{56}$$

where N_W denotes the number of originally wrongly predicted samples and $N_{W\to C}$ denotes the number of samples whose predictions change from wrong to correct after rectification.

Similarly, we define the degradation rate (DR) as the proportion of originally correctly predicted samples that become wrongly predicted after rectification:

$$\mathrm{DR} = \frac{N_{C\to W}}{N_C}, \tag{57}$$

where N_C denotes the number of originally correctly predicted samples and $N_{C\to W}$ denotes the number of samples whose predictions change from correct to wrong after rectification.

Table 17: Recovery rate (RR) and degradation rate (DR) after SGSR on five unseen deepfake datasets. Each entry reports RR / DR (%).

Detector	DFD	CDF-v2	DFDCP	DFDC	WDF	Avg.
ResNet-50	7.53 / 1.33	4.34 / 2.88	5.16 / 3.96	4.88 / 2.28	3.64 / 4.28	5.11 / 2.95
Effort	28.51 / 1.08	7.73 / 3.68	15.48 / 3.49	12.90 / 5.65	16.52 / 3.73	16.23 / 3.53
GenD-CLIP-L	37.48 / 1.07	17.21 / 2.97	25.43 / 4.40	18.07 / 6.25	29.08 / 10.95	25.45 / 5.13
BPL	18.42 / 1.06	15.56 / 1.92	18.28 / 4.18	14.07 / 4.49	13.17 / 5.21	15.90 / 3.37

Table 18: Numbers of recovered and degraded predictions after SGSR on five unseen deepfake datasets. Each entry reports the number of wrong-to-correct / correct-to-wrong predictions.

Detector	DFD	CDF-v2	DFDCP	DFDC	WDF	Total
ResNet-50	2434 / 937	369 / 228	386 / 386	2622 / 1810	339 / 705	6150 / 4066
Effort	1156 / 250	91 / 109	179 / 135	1304 / 1580	267 / 180	2997 / 2254
GenD-CLIP-L	6280 / 915	758 / 357	1061 / 574	6049 / 6226	1837 / 2132	15985 / 10204
BPL	4617 / 820	1151 / 173	988 / 494	5198 / 4318	1209 / 866	13163 / 6671

As shown in Table 17, the macro-average recovery rate is higher than the degradation rate for all four detectors. Specifically, ResNet-50, Effort, GenD-CLIP-L, and BPL recover 5.11%, 16.23%, 25.45%, and 15.90% of their originally wrong predictions, respectively, while degrading only 2.95%, 3.53%, 5.13%, and 3.37% of their originally correct predictions.

To complement the rate-based analysis, Table 18 reports the corresponding absolute transition counts. When aggregated across the five target datasets, the number of wrong-to-correct transitions exceeds the number of correct-to-wrong transitions for every detector. In particular, SGSR produces net gains of 2084, 743, 5781, and 6492 correctly classified samples for ResNet-50, Effort, GenD-CLIP-L, and BPL, respectively.

Overall, these results reveal a consistent asymmetric rectification behavior. Originally wrong predictions exhibit larger representation- and prediction-level changes than correct predictions, and the resulting prediction transitions favor recovery over degradation both in relative rates and in absolute counts. Together, these observations provide empirical evidence that SGSR acts more strongly on samples associated with incorrect decisions while preserving a larger proportion of already correct predictions.

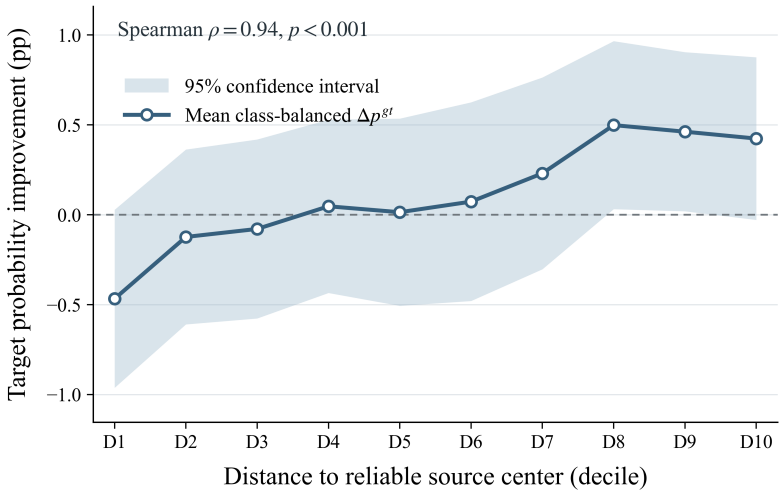


Figure 5: Class-balanced target probability improvement across deciles of increasing representation deviation. Results are aggregated over ten detectors and five unseen Deepfake datasets. The solid curve shows the mean class-balanced Δp^{gt}, and the shaded region indicates the 95% confidence interval. D1 and D10 correspond to the closest and farthest target samples from the reliable source center, respectively.

A.11 Rectification Effect across Representation Deviation

We further analyze how the effect of SGSR varies with the degree of representation deviation from the reliable source representation region. The analysis is conducted on ten representative detectors, including ResNet-50, Effort, CLIP-LoRA, CLIP-VPT, UIA-ViT, GenD, BPL, GenD-DINOv3, SBI, and Xception, across five unseen Deepfake datasets.

For each target representation $\mathbf{z}_i$, we compute its Euclidean distance to the reliable source center $\boldsymbol{\mu}$,

$$d_i = \|\mathbf{z}_i - \boldsymbol{\mu}\|_2 . \tag{58}$$

Within each detector–dataset pair, all target samples are ranked according to d_i and divided into ten equally sized deciles, denoted by D1–D10, where D1 contains the samples closest to the reliable source center and D10 contains those farthest from it.

To quantify whether rectification moves the detector prediction toward the ground-truth class, we define the target probability improvement for each sample as

$$\Delta p_i^{\mathrm{gt}} = (2y_i - 1)\left(p_i^{\mathrm{SGSR}} - p_i^{\mathrm{Base}}\right), \tag{59}$$

where $y_i \in \{0, 1\}$ is the ground-truth label, and p_i^{Base} and p_i^{SGSR} denote the predicted probabilities for the positive class before and after rectification, respectively. Thus, $\Delta p_i^{\mathrm{gt}} > 0$ indicates that rectification moves the prediction toward the ground-truth class, whereas $\Delta p_i^{\mathrm{gt}} < 0$ indicates the opposite. To reduce the influence of class imbalance, we compute the class-balanced mean of Δp_i^{gt} within each decile.

Figure 5 shows a clear increasing trend in the benefit of SGSR as the representation deviation becomes larger. For samples close to the reliable source center (D1–D3), the average target probability improvement is slightly negative, indicating that these samples already lie in relatively reliable representation regions and therefore offer limited room for beneficial correction. As the distance to the reliable source center increases, the improvement gradually becomes positive and rises substantially for the distant deciles (D7–D10), peaking around D8.

Quantitatively, the Spearman rank correlation between the decile index and the mean class-balanced target probability improvement is

$$\rho_{\mathrm{Spearman}} = 0.939, \qquad p = 5.48 \times 10^{-5}, \tag{60}$$

indicating a strong and statistically significant monotonic relationship. These results show that the beneficial prediction changes induced by SGSR become more pronounced for samples with larger representation deviations. This observation is consistent with the empirical rectification behavior of SGSR, where more deviated samples tend to exhibit larger beneficial prediction changes.

A.12 Qualitative Visualization of Representation Rectification

To qualitatively examine how SGSR modifies target representations, we visualize the representation distributions before and after rectification using PCA. We consider five representative CLIP adaptation strategies, including VPT, full fine-tuning (FFT), LoRA, LayerNorm tuning (LN), and BitFit. For each model, the reliable source representations, original target representations, and rectified target representations are jointly projected onto the first two principal components.

For visualization, we use DFDCP as the target dataset for Deepfake detection and `glide_50_27` as the target dataset for synthetic image detection. As shown in Fig. 6, across different adaptation strategies and representation depths, the rectified target representations generally move toward regions better supported by the reliable source distribution. Importantly, the rectified representations do not simply collapse toward the source center. Instead, they retain substantial variation in the target representation while shifting toward the source-characterized reliable region. This behavior is consistent with the formulation of SGSR, which performs direction-dependent rectification according to the estimated reliable representation geometry rather than uniformly pulling all representations toward a single center.

This qualitative behavior can be observed consistently across the analyzed layers and adaptation strategies. In many cases, the rectified target representations become better aligned with the support of the reliable source distribution, while still preserving their original structural diversity. Overall, these visualizations provide qualitative evidence that SGSR rectifies shifted target representations in a structured manner rather than performing a trivial global shrinkage.

A.13 Analysis of Larger Gains on Synthetic Image Detection

The main experiments reveal an interesting difference between the two detection scenarios: SGSR yields a larger average improvement on synthetic image detection than on Deepfake detection. Specifically, the average AUC improvement is $+0.80$ points on the five unseen Deepfake datasets, compared with $+4.08$ points on the UniversalFakeDetect (UFD) benchmark. We investigate this difference from the perspective of representation deviation and the resulting rectification magnitude.

Source-standardized representation deviation. Directly comparing Euclidean representation distances across different fine-tuning strategies can be affected by differences in representation scale and covariance. We therefore measure target representation deviation relative to the second-order statistics of reliable source representations.

For each intermediate layer ℓ, we estimate the reliable source mean $\boldsymbol{\mu}_\ell$ and covariance $\widehat{\boldsymbol{\Sigma}}_\ell$ using correctly predicted source training representations. We use the same low-rank covariance approximation as in SGSR,

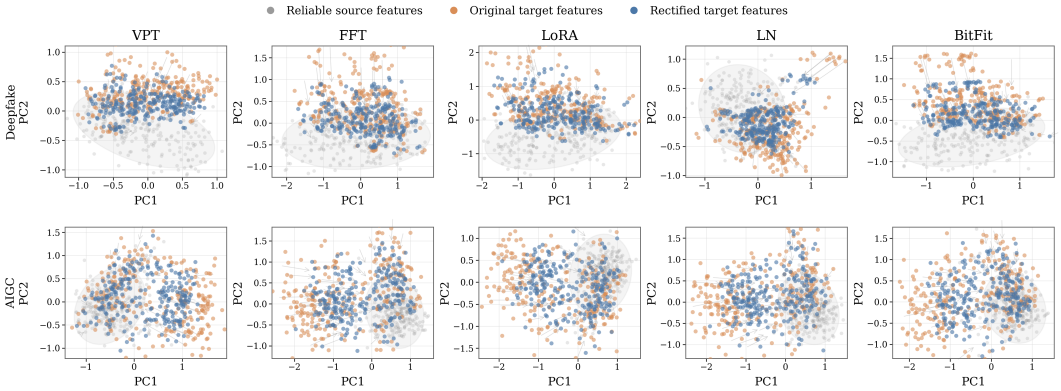
$$\widehat{\boldsymbol{\Sigma}}_\ell = \mathbf{U}_\ell \boldsymbol{\Lambda}_\ell \mathbf{U}_\ell^\top + \sigma_{\perp,\ell}^2 \left(\mathbf{I} - \mathbf{U}_\ell \mathbf{U}_\ell^\top \right). \tag{61}$$

For a baseline target representation $\mathbf{z}_{i,\ell}$, we compute its dimension-normalized covariance-aware deviation as

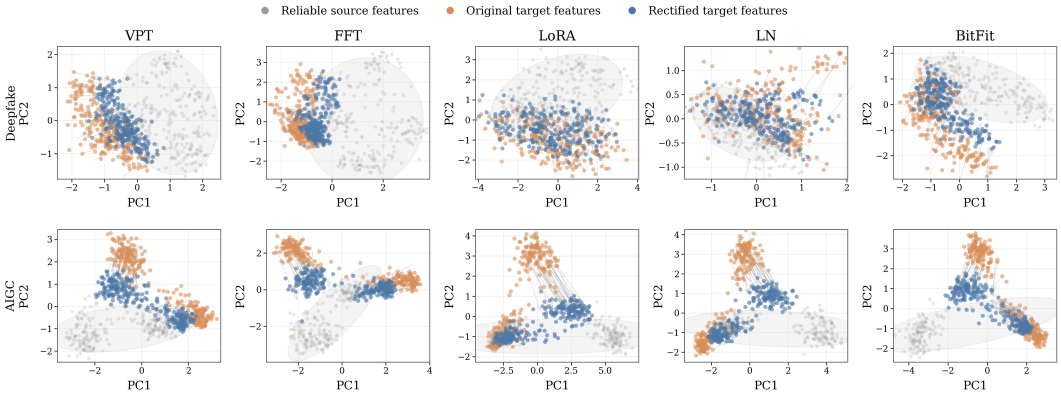
$$q_{i,\ell} = \frac{1}{d} \left(\mathbf{z}_{i,\ell} - \boldsymbol{\mu}_\ell \right)^\top \widehat{\boldsymbol{\Sigma}}_\ell^{-1} \left(\mathbf{z}_{i,\ell} - \boldsymbol{\mu}_\ell \right). \tag{62}$$

To further account for scale differences across fine-tuning strategies and layers, we normalize the target deviation using held-out correctly predicted source representations:

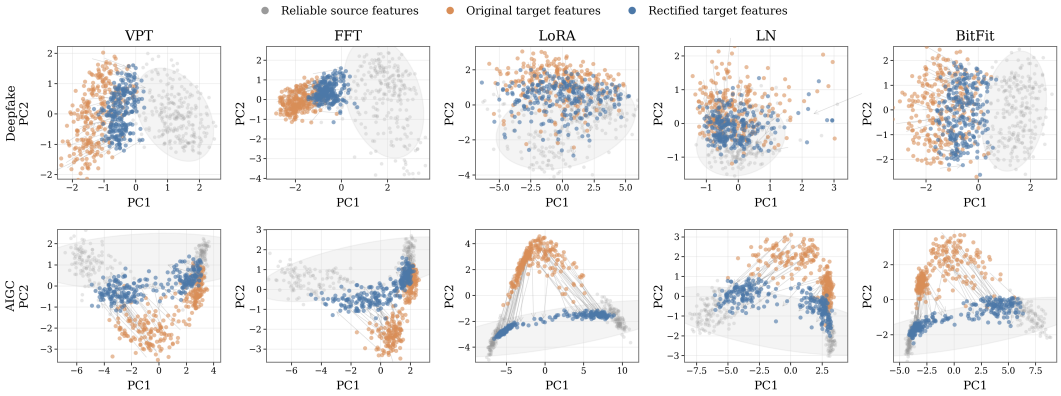
$$\mathrm{SSD}_{m,t,\ell} = \frac{\mathbb{E}_{\mathbf{z} \sim \mathcal{T}_t} \left[q_{i,\ell} \right]}{\mathbb{E}_{\mathbf{z} \sim \mathcal{H}_{\mathrm{source,val}}} \left[q_{i,\ell} \right]}, \tag{63}$$



(a) Layer 6.



(b) Layer 12.



(c) Layer 18.

Figure 6: **Qualitative visualization of representation rectification across representation depths.** PCA projections are shown for five CLIP adaptation strategies: VPT, FFT, LoRA, LN, and BitFit, at layers 6, 12, and 18. Within each panel, the top row corresponds to Deepfake detection on DFDCP, and the bottom row corresponds to synthetic image detection on `glide_50_27`. Gray, orange, and blue points denote reliable source representations, original target representations, and rectified target representations, respectively. Lines connect each original target representation to its rectified counterpart.

where m denotes the fine-tuning strategy and t denotes the target dataset. Let L denote the number of intermediate layers in the detector. We then average over all intermediate layers:

$$\mathrm{SSD}_{m,t} = \frac{1}{L}\sum_{\ell=1}^{L}\mathrm{SSD}_{m,t,\ell}. \tag{64}$$

Table 19: Comparison of source-standardized deviation (SSD) and representation rectification magnitude (M) between Deepfake and synthetic image detection. Synthetic image results are evaluated on the UFD benchmark. The UFD/DF columns report the ratio between the two task settings.

Fine-tuning	SSD_{DF}	SSD_{Syn}	UFD/DF	M_{DF}	M_{Syn}	UFD/DF
VPT	2.106	2.700	1.282×	1.739	2.935	1.688×
FFT	2.164	3.148	1.455×	2.362	3.352	1.419×
LoRA	1.741	3.091	1.775×	2.512	4.891	1.947×
LN	2.007	3.125	1.557×	1.056	3.633	3.441×
BitFit	2.125	3.178	1.496×	1.972	3.667	1.860×
Macro Avg.	**2.029**	**3.049**	**1.503×**	**1.928**	**3.696**	**1.917×**

A larger Source-standardized deviation (SSD) indicates that target representations deviate more strongly from the representation geometry characterized by reliable source samples.

Rectification magnitude. We next directly measure the magnitude of the representation modification introduced by SGSR. For each intermediate layer, we compute

$$M_{m,t,\ell} = \frac{1}{N_t} \sum_{i=1}^{N_t} \left\| \hat{\mathbf{u}}_{i,\ell} - \mathbf{z}_{i,\ell} \right\|_2 , \tag{65}$$

where $\mathbf{z}_{i,\ell}$ is the baseline representation and $\hat{\mathbf{u}}_{i,\ell}$ is the representation after rectification. The layer-averaged magnitude is

$$M_{m,t} = \frac{1}{L} \sum_{\ell=1}^{L} M_{m,t,\ell}. \tag{66}$$

Since all compared settings use the same CLIP backbone, token representation, and representation dimensionality, the resulting representation-space magnitudes are reported as a descriptive comparison.

Analysis. As shown in Table 19, synthetic-image representations consistently exhibit larger source-standardized deviations than Deepfake representations across all five fine-tuning strategies. On average, SSD increases from 2.029 for Deepfake detection to 3.049 for synthetic image detection, corresponding to a $1.503\times$ larger deviation relative to the reliable source representation geometry. This suggests that unseen generators in UFD induce more substantial representation shifts from the reliable states characterized by correctly predicted source samples.

More importantly, the larger representation deviation is accompanied by a substantially larger actual rectification magnitude. The average representation modification increases from 1.928 on Deepfake datasets to 3.696 on synthetic image datasets, corresponding to a $1.917\times$ increase. This behavior is consistent across all five fine-tuning strategies, indicating that SGSR intervenes more strongly on synthetic image representations.

These results are consistent with the rectification behavior analyzed in Sec. A.10, where larger representation deviations are shown to induce stronger representation-level and prediction-level responses under SGSR. Here, synthetic image representations exhibit both larger source-standardized deviations and larger rectification magnitudes than Deepfake representations, suggesting that unseen synthetic image generators introduce more substantial rectifiable representation shifts.

This difference is also reflected in the final detection performance. As reported in the main experiments, SGSR improves the average AUC by $+0.80$ points on the five unseen Deepfake datasets, compared with $+4.08$ points on UFD. Together, these results provide representation-level evidence consistent with the larger performance gains observed in synthetic image detection.