TOWARDS GENERALIZABLE DEEPFAKE DETECTION VIA REAL DISTRIBUTION BIAS CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

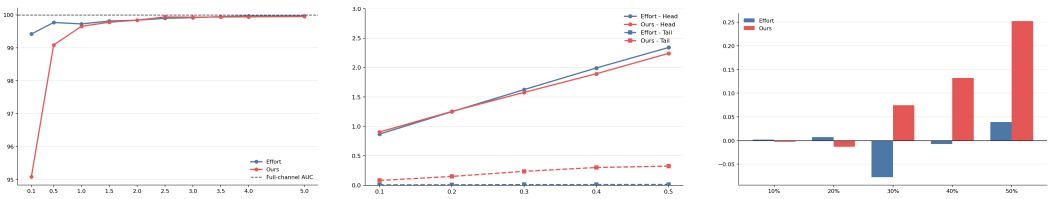
Deepfake detection can benefit from the real-image knowledge encoded by pretrained foundation models. Existing approaches seek to retain this knowledge by preserving a subset of pretrained parameters while updating the remaining parameters for deepfake detection. However, freezing pretrained weights cannot freeze their output responses, which can cause channel-parameter mismatch in downstream layers and undermine the use of pretrained knowledge. In principle, gradient-based optimization may mitigate this mismatch, but our analysis reveals that classification supervision overemphasizes a small set of head channels while leaving tail channels weakly constrained. Consequently, the misaligned tail responses can interfere with predictions in unseen domains and hinder generalization. To address this problem, we propose the Real Distribution Bias Correction (RDBC) framework, which uses real-image statistics as reference signals to calibrate the under-constrained channel-response space. Specifically, the Real-Kurtosis-Guided Channel Alignment (RKA) module uses the input-side kurtosis of real samples as a reference to align the output cross-channel response structure at the final frozen normalization layer, addressing accumulated response drift and constraining weakly supervised tail-channel responses. The Real-Covariance-Guided Channel Enhancement (RCE) module uses real-feature covariance to propagate the centered deviations of highly responsive channels to other co-varying channels, thereby reducing the reliance on isolated head-channel responses at shallow layers. RCE further samples its statistics from a Gaussian-fitted statistical neighborhood, expanding the coverage of the real-feature reference. Through the complementary calibration of shallow and deep channel responses, RDBC corrects the channel-parameter mismatch while enhancing real-fake separability. Extensive experiments demonstrate that RDBC achieves state-of-the-art performance in cross-domain deepfake detection.

1 INTRODUCTION

Advanced image generation models have significantly reduced the barriers to synthesizing realistic images, leading to the proliferation of malicious content such as deepfakes Wang et al. (2024). To mitigate this risk, many deepfake detection methods have been proposed to detect obvious artifacts in manipulated images Yan et al. (2025b); Wang et al. (2026a); Tan et al. (2026b). These approaches achieve strong performance when trained and tested on datasets from the same domain. However, they are prone to overfitting to salient artifacts in the training set and continue to exhibit poor generalization, with their accuracy dropping sharply on unseen forgeries Han et al. (2025).

To alleviate this problem, most previous studies employ data augmentation to expand the source-domain forgeries to approximate those in the target domain Sun et al. (2024); Yan et al. (2024). Although considerable progress has been made Wang & Deng (2021); Cheng et al. (2023), these approaches are built on an implicit assumption: *forgery patterns can be predicted from existing observations*. However, as deepfake techniques evolve rapidly Esser et al. (2024), forgeries have become diverse and unpredictable, rendering this assumption invalid. Unlike detection models that search for local artifacts, humans distinguish real from fake data by additionally evaluating the consistency between target data and prior cognitive patterns of real data (*i.e.*, pattern consistency). Although imaging technologies continuously evolve and exhibit diverse sensor characteristics, their core mechanism remains a high-fidelity mapping of physical-world regularities. This highlights a

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.



(a) AUC Recovery of High-Contribution Channels (b) Head-Tail Allocation of Feature Gradients (c) Interference from Low-Contribution Channels

Figure 1: Head-channel dominance and tail-channel interference. Results aggregate **all of the randomly selected images** rather than separate examples: (a) AUC recovery in the source domain, (b) gradient share in the source domain, (c) signed contribution in the target domain.

Table 1: Shallow CLIP features exhibit substantially higher cross-channel kurtosis and response concentration than EfficientNet features. The results are obtained by evaluating on the FF++ dataset.

Backbone	Block	D	Pearson Kurtosis		Response Concentration	
			Real	Fake	Real	Fake
CLIP	1	1024	846.01 ± 4.26	845.65 ± 4.10	0.826	0.826
	4	1024	815.12 ± 5.68	815.38 ± 5.62	0.796	0.796
	8	1024	603.35 ± 32.45	603.60 ± 32.51	0.589	0.589
EfficientNet	1	24	2.57 ± 0.58	2.50 ± 0.52	0.107	0.104
	4	32	2.87 ± 0.68	2.70 ± 0.50	0.090	0.084
	8	56	3.15 ± 0.47	3.32 ± 0.51	0.056	0.059

critical insight: compared with the continuously evolving manipulated data, the real data is inherently more deterministic and robust to modeling Nadarajah (2009); Simoncelli & Olshausen (2001).

Pretrained foundation models encode broad knowledge of real images and offer a useful starting point for deepfake detection. To retain this knowledge during fine-tuning, approaches such as Effort (Yan et al., 2025a; Ji et al., 2026; Lek et al., 2026; Sun et al., 2026; Yang et al., 2026) preserve part of the pretrained parameters while updating the remaining parameters. However, freezing pretrained weights cannot freeze their output responses. As the trainable parameters change, shifted features enter downstream frozen mappings, which can cause channel-parameter mismatch and undermine the use of pretrained knowledge. Gradient-based optimization may mitigate this mismatch, but whether classification supervision sufficiently calibrates the full channel space remains unclear.

We analyze this problem at the channel level. As shown in Figure 1(a), retaining only the top 5% of channels ranked by source-domain logit contribution recovers more than 95% of the full AUC. Figure 1(b) measures the feature gradients of the same ranked channels: high-contribution channels receive a disproportionately large gradient share, whereas low-contribution tail channels receive very little. Together, these results show that channels contributing little to source-domain classification also receive limited direct classification gradients, leaving little optimization pressure to align their responses with frozen mappings. Figure 1(c) provides evidence of the resulting interference: on target-domain images, the accumulated signed contributions of these low-contribution channels can oppose the ground-truth class and affect predictions. Head-channel dominance also raises concerns about feature sparsity and shortcut learning. Here, we focus on its effect on the remaining channels: weakly supervised tail responses can remain mismatched with frozen mappings and interfere with target-domain discrimination. We therefore introduce the Real Distribution Bias Correction (RDBC) framework, which uses real-image statistics to calibrate the channel-response space.

As shown in Figure 2, RDBC comprises two complementary modules. **i) Real-Kurtosis-Guided Channel Alignment (RKA).** RKA addresses the accumulated channel-response mismatch at the frozen post-layer normalization (post-LN) layer preceding the classifier. For real samples, it uses the kurtosis of the post-LN input features to supervise the kurtosis of its output features. This constrains the cross-channel response structure fed to the classifier, including weakly supervised tail channels, and directly mitigates the effect of the mismatch on classification. **ii) Real-Covariance-Guided Channel Enhancement (RCE).** Compared with conventional CNN detectors, CLIP exhibits extremely high cross-channel kurtosis in its shallow features, as shown in Table 1. This indicates that a few shallow channels produce extreme responses, motivating channel calibration at the source of the

representation. RCE uses real-feature covariance to propagate the centered deviations of highly responsive channels to other co-varying channels, promoting a more balanced channel-response structure and reducing reliance on isolated head-channel responses. Meanwhile, to reduce overfitting to source-domain real features, RCE samples statistics from a Gaussian-fitted statistical neighborhood, thereby broadening its real-feature reference. Our main contributions are summarized as follows:

- We identify channel-parameter mismatch in parameter-preserving fine-tuning and show that classification supervision concentrates on head channels while tail channels remain weakly constrained and can interfere with unseen-domain predictions.
- We propose the Real Distribution Bias Correction framework, combining deep kurtosis-guided alignment to calibrate accumulated channel-response mismatch, with shallow real-covariance-guided enhancement to mitigate extreme-response concentration.
- Extensive experiments across multiple deepfake benchmarks validate the individual and complementary effectiveness of RKA and RCE, while demonstrating the strong generalization performance of RDBC across unseen target domains.

2 Related work

Pretrained representations for deepfake detection. Large-scale visual pretraining provides transferable knowledge that can reduce the dependence on specific artifacts. Ojha et al. (Ojha et al., 2023) demonstrate that fixed representations from CLIP (Radford et al., 2021), combined with nearest-neighbor classification, generalize effectively to unseen generators. Subsequent studies adapt foundation models using facial-component-guided side networks (Han et al., 2025) or spatiotemporal adapters combined with video-level blending (Yan et al., 2025b). Effort (Yan et al., 2025a) further decomposes pretrained weights through singular value decomposition, preserving the principal components while adapting an orthogonal residual subspace. These approaches preserve pretrained knowledge primarily through feature reuse or parameter-level constraints. However, frozen parameters do not guarantee that their output responses remain statistically compatible with features shifted by task adaptation. In contrast, our RDBC framework uses real-feature statistics to calibrate the complete channel space and maintain its compatibility with frozen pretrained mappings.

Feature statistics for domain generalization. Feature statistics provide an effective means of synthesizing domain variations without access to target-domain data Ding et al. (2026); Tan et al. (2026a); Li et al. (2026). MixStyle (Zhou et al., 2021) linearly mixes the instance-level means and standard deviations to synthesize new styles. Going beyond pairwise interpolation, DSU (Li et al., 2022) estimates the uncertainty of channel-wise means and standard deviations and samples new statistics to simulate potential domain shifts. CSU (Zhang et al., 2024) further observes that channel-wise style statistics are correlated and uses covariance-based sampling to preserve their inter-channel dependencies. Shifting the focus toward higher-order statistical characteristics, our RDBC uses statistics derived from source-domain real features as calibration references for deepfake detection(Shahjahan et al., 2026; Kotyan et al., 2026). RKA employs cross-channel kurtosis to align the response structure across the frozen layer, while RCE uses covariance to enhance co-varying channel responses and samples from a statistical neighborhood to broaden the real-feature reference.

3 Method

3.1 Channel-parameter mismatch and weak tail-channel constraints

Let $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^{N}$ denote the source-domain training set, where $y_i \in \{0, 1\}$ is the real-fake label. Let W_{fix} and W_{tune} denote the frozen and trainable parameter components of the pretrained vision foundation model, respectively. When all parameters retain their pretrained values, the feature x_{l-1} entering layer l is produced by the pretrained upstream network. In this state, x_{l-1} and the frozen mapping f_{fix} retain the input–mapping relationship established during pretraining, yielding

$$x_l = f_{\text{fix}}(x_{l-1}; W_{\text{fix}}). \tag{1}$$

During fine-tuning, updates to W_{tune} in the preceding layers change the input to the same frozen mapping f_{fix} from x_{l-1} to x'_{l-1}:

$$x'_{l-1} = x_{l-1} + \delta_{l-1}, \tag{2}$$

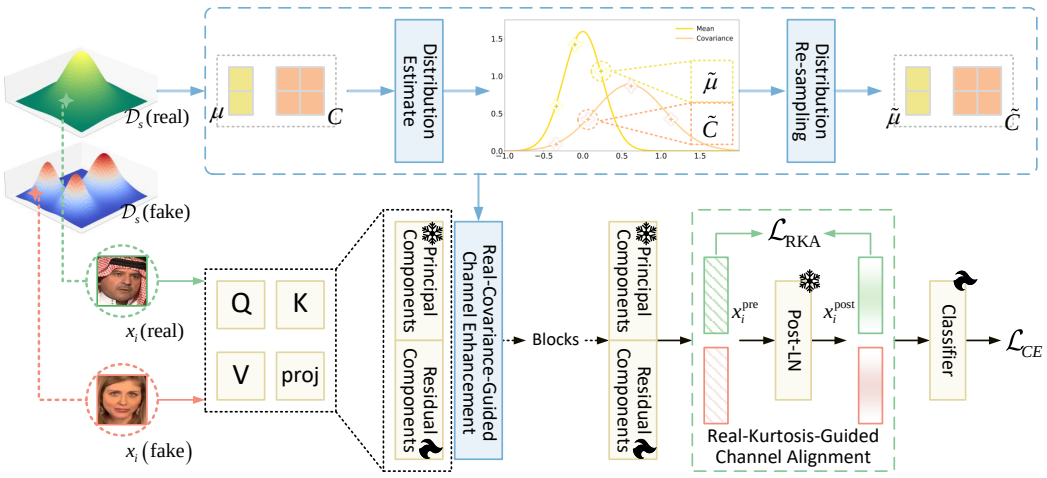


Figure 2: The proposed Real Distribution Bias Correction framework, comprising Real-Kurtosis-Guided Channel Alignment (RKA) for aligning high-order statistic structure, and Real-Covariance-Guided Channel Enhancement (RCE) for enhancing real-covariance-supported responses.

where δ_{l-1} is the feature change induced by fine-tuning. Because W_{fix} remains unchanged while f_{fix} receives x'_{l-1} rather than x_{l-1}, the resulting output change is

$$\Delta x_l = f_{\text{fix}}(x_{l-1} + \delta_{l-1}; W_{\text{fix}}) - f_{\text{fix}}(x_{l-1}; W_{\text{fix}}). \tag{3}$$

Thus, freezing W_{fix} preserves the pretrained mapping but does not freeze its channel responses. When the shifted input x'_{l-1} is no longer compatible with f_{fix}, we refer to the resulting drift as *channel-parameter mismatch*. This drift can accumulate across downstream layers and alter the representation fed to the classifier. Gradient-based optimization may mitigate this mismatch by updating W_{tune}. However, Figure 1 shows that classification supervision can achieve source-domain discrimination through a small set of high-contribution head channels, while the same low-contribution tail channels receive little direct cross-entropy feature gradient. The mismatch in these weakly constrained tail responses may therefore persist despite successful source-domain classification.

Because frozen mappings are retained to preserve pretrained real-image knowledge, our Real Distribution Bias Correction (RDBC) framework uses real-image statistics to calibrate their responses while allowing the trainable component to learn forgery-discriminative responses. As illustrated in Figure 2, RKA uses the input-side kurtosis of real samples to guide the output response structure of the frozen post-layer normalization, where accumulated mismatch directly affects the classifier. RCE acts on shallow representations to reduce reliance on isolated high-response channels.

3.2 Real-Kurtosis-Guided Channel Alignment

The post-LN maps the output of the last visual block to the representation used by the classifier. Let $x_i^{\text{pre}} \in \mathbb{R}^D$ and $x_i^{\text{post}} \in \mathbb{R}^D$ denote its input and output representations, respectively:

$$x_{ij}^{\text{post}} = \gamma_j \left[\text{Norm}(x_i^{\text{pre}})\right]_j + \beta_j, \tag{4}$$

where $\text{Norm}(\cdot)$ denotes normalization without the affine transformation, and the pretrained channel-wise affine parameters γ_j and β_j remain frozen during fine-tuning. As discussed in Section 3.1, upstream fine-tuning can shift x_i^{pre} while these parameters remain fixed. The resulting channel-parameter mismatch changes the response structure of x_i^{post} at the classifier input. To examine its effect on classification, let W_c and b_c denote the classifier weight vector and bias for class c. The correct-class logit margin is

$$M_i = (W_{y_i} - W_{1-y_i})^\top x_i^{\text{post}} + b_{y_i} - b_{1-y_i}. \tag{5}$$

Suppose upstream fine-tuning changes the input from x_i^{pre} to $x_i^{\text{pre}\prime}$. Let $\delta_i = \text{Norm}(x_i^{\text{pre}\prime}) - \text{Norm}(x_i^{\text{pre}})$ denote the corresponding change in the normalized response. Holding the classifier

parameters fixed to isolate this response change, the change in the correct-class margin is

$$\Delta M_i = \sum_{j=1}^{D} (W_{y_i,j} - W_{1-y_i,j}) \gamma_j \delta_{ij}. \tag{6}$$

Eq. (6) shows that the response change δ_{ij} of channel j contributes to the correct-class margin change ΔM_i through its corresponding frozen parameter γ_j and classifier weight $W_{y_i,j} - W_{1-y_i,j}$. It thus shows how channel-parameter mismatch can affect classification: if a shifted response cannot be compatible with these parameters, it may reduce the correct-class margin. Because ΔM_i sums over all channels, such contributions can accumulate across tail channels. Figure 1 shows that the low-contribution tail channels receive little cross-entropy feature gradient supervision, and contribute against the correct class in the target domain. These observations motivate constraining the overall post-LN response structure, including the weakly supervised tail channels.

RKA uses the post-LN input x_i^{pre} of each real sample as a reference to calibrate its corresponding output x_i^{post}. Specifically, it captures the concentration of extreme responses across the D channels using sample-wise Pearson kurtosis:

$$\kappa(x_i) = \frac{\frac{1}{D}\sum_{j=1}^{D}(x_{ij} - \bar{x}_i)^4}{\left(\frac{1}{D}\sum_{j=1}^{D}(x_{ij} - \bar{x}_i)^2\right)^2}, \qquad \bar{x}_i = \frac{1}{D}\sum_{j=1}^{D} x_{ij}. \tag{7}$$

For each real sample, we compute the cross-channel kurtosis of x_i^{pre} and x_i^{post}, denoted by κ_i^{pre} and κ_i^{post}. RKA then aligns the output kurtosis with its input-side reference through

$$\mathcal{L}_{\mathrm{RKA}} = \frac{1}{|\mathcal{B}_r|} \sum_{i \in \mathcal{B}_r} \mathrm{SmoothL1}\left(\log \kappa_i^{\mathrm{post}}, \log \mathrm{detach}(\kappa_i^{\mathrm{pre}})\right), \tag{8}$$

where $\mathcal{B}_r$ denotes the real samples in the current batch and $\mathrm{detach}(\cdot)$ stops gradients through the input-side reference. Using the current last-block response of each real sample as a sample-adaptive reference, this loss limits changes in cross-channel kurtosis across the frozen post-LN. It thereby provides a distribution-level constraint on the full output channel vector, including channels weakly supervised by classification. During training, fake samples are not directly subject to this alignment loss and can develop task-specific discriminative responses.

3.3 Real-Covariance-Guided Channel Enhancement

Table 1 shows that shallow pretrained CLIP features exhibit substantially higher cross-channel kurtosis than conventional CNN features, indicating that a few channels produce extreme responses early in feature extraction. Together with the head-channel dominance observed in Figure 1, this motivates RCE to calibrate shallow channel responses. RCE uses source-domain real covariance to propagate centered deviations from highly responsive channels to other channels that co-vary with them, promoting a more balanced response structure.

Let x_{ip} denote the p-th patch feature of image x_i. We rank the patches by response magnitude $\|x_{ip}\|_2$. For a selection ratio $\rho \in (0, 1]$, let $\mathcal{P}_i$ contain the indices of the top ρ fraction of patches. The selected patches of each source-domain real image are averaged into one image-level feature:

$$x_i^r = \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} x_{ip}. \tag{9}$$

From the resulting N_r high-response averaged real features, we calculate the mean and covariance:

$$\mu = \frac{1}{N_r}\sum_{i=1}^{N_r} x_i^r, \quad C = \frac{1}{N_r - 1}\sum_{i=1}^{N_r}(x_i^r - \mu)(x_i^r - \mu)^\top, \tag{10}$$

Covariance-guided channel enhancement. Using the statistics in Eq. (10), we apply the same covariance-guided transformation to selected high-response patches from both real and fake images. For channel j of a selected patch x_{ip}, the transformation is

$$\hat{x}_{ip,j} = x_{ip,j} + \sum_{k=1}^{D} C_{jk}(x_{ip,k} - \mu_k), \tag{11}$$

where subtracting μ expresses the deviation from the source-domain real response center, and multiplication by C transfers that deviation according to channel dependencies. In particular, a deviation in a highly responsive channel k contributes to channel j when $C_{jk} \neq 0$. This encourages statistically related channels to participate in the representation instead of relying on an isolated response. During backpropagation, the gradients can be transmitted between co-varying channels, providing an additional path for weakly used channels to participate in optimization, as derived in Appendix C.

Further, because C is positive semidefinite ($C \succeq 0$), the covariance-guided transformation can enlarge the real-fake mean difference among selected patches, as detailed in Appendix A:

$$\hat{d} = (I + C)d, \qquad \|\hat{d}\|_2 \geq \|d\|_2, \tag{12}$$

where μ_r and μ_f are the means of the selected real and fake patch features. $d = \mu_f - \mu_r$, and $\hat{d}$ is defined analogously after applying the same transformation to both classes. I is the identity matrix.

Gaussian-fitted source-real statistical extrapolation. Using only the fixed source-real pair (μ, C) in Eq. (11) throughout training can reinforce overfitting to the source domain. To broaden the real-feature reference, we draw B subsets of n image-level real features x_i^r and compute a mean-covariance pair (μ_b, C_b) for each subset. Motivated by the asymptotic normality of subset statistics, we fit factorized Gaussian distributions to the subset means and the log directional variances of the subset covariances. For covariance sampling, we use the fixed eigenspace Q of C and reconstruct the sampled matrix with positive eigenvalues to preserve positive definiteness; details are given in Appendix A. During training, we sample

$$\begin{aligned} \widetilde{\mu} &= \bar{\mu} + \sigma_\mu \odot z_\mu, \\ \widetilde{C} &= Q \,\mathrm{diag}[\exp(\bar{s} + \sigma_s \odot z_C)]\, Q^\top, \end{aligned} \tag{13}$$

where $\bar{\mu}$ and σ_μ are the element-wise mean and standard deviation of $\{\mu_b\}_{b=1}^B$, while $\bar{s}$ and σ_s are those of the log directional variances derived from $\{C_b\}_{b=1}^B$. The vectors z_μ and z_C are independent standard normal samples. By re-sampling statistics pairs $(\widetilde{\mu}, \widetilde{C})$, these fitted distributions provide a continuous neighborhood of source-real statistics beyond the finite subset estimates.

At the beginning of each epoch, one pair $(\widetilde{\mu}, \widetilde{C})$ is sampled and substituted into Eq. (11) for both real and fake samples. Because the same pair is applied to both classes and $\widetilde{C}$ is positive definite, the mean-gap relation in Eq. (12) also holds with $\widetilde{C}$. At inference, sampling is disabled, and a deterministic pair is formed from the centers of the fitted latent distributions $(\bar{\mu}, Q\,\mathrm{diag}(e^{\bar{s}_1}, \ldots, e^{\bar{s}_D})Q^\top)$.

3.4 Training and inference

The classification loss $\mathcal{L}_{\mathrm{CE}}$ is computed over all real and fake samples, whereas the RKA objective $\mathcal{L}_{\mathrm{RKA}}$ is computed only over real samples. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda_{\mathrm{RKA}} \mathcal{L}_{\mathrm{RKA}}, \tag{14}$$

where λ_{RKA} controls the strength of RKA alignment and is set to 0.01.

4 Experiments and evaluation protocol

Input and training protocol. We train our model on the raw frames of FaceForensics++ (FF++) Rössler et al. (2019), with real frames from the original videos and fake frames from Deepfakes, Face2Face, FaceSwap, and NeuralTextures. Celeb-DF Li et al. (2020), DFDC Dolhansky et al. (2020), DFDCp, and UADFV Yang et al. (2019) are used in both frame-level and video-level target-domain evaluations. The frame-level comparison additionally includes DiFF and DFD, while the video-level comparison additionally includes WDF and FFIW. We adopt the network architecture of Effort Yan et al. (2025a), built on a pretrained CLIP ViT-L/14 visual encoder, as our backbone. We use the same source data and protocol as the backbone, and we do not use the additional loss terms introduced in Effort. The detected faces are cropped using the landmarks and face boxes, resized to 380×380. Each batch contains eight real and eight fake images. Both models follow the same 11-epoch training schedule. We train on a single NVIDIA GeForce RTX 3090 with Adam, a learning rate of 2×10^{-4}, weight decay of 5×10^{-4}, and random seed 1024.

Table 2: Video-level AUC comparison (%). The best performance is marked in bold. ‡: We re-implement the detector. -: The relevant results are not reported in the original paper.

Method	Venue	Testing Dataset						
		FF++	Celeb-DF	DFDC	DFDCp	UADFV	WDF	FFIW
‡CORE Ni et al. (2022)	CVPRW'22	96.61	79.45	62.60	75.74	65.41	72.4	71.0
‡RECCE Cao et al. (2022)	CVPR'22	96.95	69.71	62.82	74.19	78.61	75.6	71.1
‡SBI* Shiohara & Yamasaki (2022)	CVPR'22	85.16	93.18	72.42	86.15	97.28	70.3	86.6
‡UCF Yan et al. (2023)	ICCV'23	97.16	81.90	66.21	80.94	97.15	77.4	69.7
FoCus Tian et al. (2024)	TIFS'24	99.15	76.13	68.42	76.62	-	-	-
GRU Choi et al. (2024)	CVPR'24	98.40	89.00	-	-	-	-	-
ProDet Cheng et al. (2024)	NeurIPS'24	95.91	84.48	72.40	81.16	-	78.1	75.1
‡FreqBlender* Li et al. (2024)	NeurIPS'24	93.29	92.65	73.15	84.56	94.79	-	-
‡Effort Yan et al. (2025a)	ICML'25	98.85	93.68	74.49	86.22	98.07	84.8	92.1
‡VLFFD Sun et al. (2025)	CVPR'25	98.64	89.80	71.80	84.74	96.71	-	-
SRI-Net Fei et al. (2026)	ICLR'26	-	95.5	-	-	-	-	-
SD^2 Jung et al. (2026)	ICLR'26	97.52	-	80.36	-	96.44	82.80	-
RDBC (Ours, video-level)	-	**99.41**	**95.93**	**81.11**	**88.27**	**98.72**	**85.03**	**93.52**

Table 3: Frame-level AUC comparison (%). The best performance is marked in bold. ‡: We re-implement the detector. -: The relevant results are not reported in the original paper.

Method	Venue	Testing Dataset						
		FF++	Celeb-DF	DFDC	DFDCp	UADFV	DiFF	DFD
SBI Shiohara & Yamasaki (2022)	CVPR'22	-	81.3	-	-	-	78.6	77.4
FIA-USA Ma et al. (2025)	NeurIPS'25	-	86.7	-	-	-	83.3	82.1
‡Effort Yan et al. (2025a)	ICML'25	97.79	85.54	71.12	75.27	95.42	82.30	86.19
‡SRI-Net Fei et al. (2026)	ICLR'26	-	87.5	-	-	-	**86.8**	89.3
RDBC (Ours, frame-level)	-	**98.11**	**88.10**	**72.95**	**76.31**	**95.64**	84.03	**89.61**

RCE statistics. RCE selects the top $\rho = 0.1$ fraction of non-CLS patches by ℓ_2 response magnitude, and only uses these training-real features to construct the statistical reference. There are $B = 50$ subsets; each contains $n = 128$ real-image features for statistical estimation. At the beginning of each training epoch, we update the reference from source-training real features and sample one mean-covariance pair shared by real and fake inputs. At inference, RCE uses the deterministic center of the fitted statistical distributions without target-domain statistics or test-time sampling.

4.1 Performance Comparison

We present the comparison results between our proposed method and SoTA models in Tables 3 and 2. The results come from three sources: values reported in the original papers, values reproduced in other papers, and values obtained from our own reimplementations (marked with ‡). We retain the precision available from each source, so the number of decimal places varies across table entries. All of them are trained on FF++. The frame-level and video-level comparisons each cover six target-domain datasets. RDBC and the baseline (Effort) follow the same evaluation protocol. We report frame-level and video-level Area Under the Curve (AUC) separately for each dataset.

Tables 3 and 2 show strong performance on the target domains. On DFDC, our approach achieves AUC scores of 72.95% at the frame level and 81.11% at the video level, surpassing Effort by 1.83 and 6.62 percentage points, respectively. On DFDCp, the corresponding gains over Effort are 1.04 and 2.05 percentage points. The particularly large video-level gain on DFDC and the consistent gains on DFDCp suggest that RDBC benefits challenging target data with quality degradation and a larger gap from the raw FF++ source domain. Under such shifts, fewer source-trained head-channel cues may remain reliable, making calibration of the broader channel-response space more valuable.

Some methods focus on enhancing generalization by data augmentation of real samples (marked with '*' in Table 2), but their performance still falls short of RDBC. These approaches introduce additional prior knowledge to simulate forgery traces, whereas RDBC uses statistics derived from real images without introducing additional priors. On Celeb-DF, RDBC improves the video-level AUC over Effort from 93.68% to 95.93%, suggesting more effective use of pretrained knowledge.

Table 4: Robustness evaluation (AUC %) against image corruptions.

Method	Clean	Compression		Blur	Gaussian Noise
		$Q = 70$	$Q = 50$	$\sigma = 3$	$\sigma^2 = 0.01$
EfficientNet	64.59	58.24	52.18	54.33	51.87
CADDM	91.63	82.15	73.40	76.82	70.55
SBI	93.18	84.62	75.31	78.45	76.20
Effort	93.68	87.35	80.12	82.64	79.88
RDBC	95.93	93.41	88.75	90.22	87.64

Table 5: AUC (%) comparison of different modules on three target datasets.

Backbone	RKA	RCE*	RCE	Testing Set		
				Celeb-DF	DFDC	DFDCp
✓				93.68	74.49	86.22
✓	✓			94.50	75.81	87.08
✓		✓		94.24	78.35	86.93
✓			✓	95.32	80.27	87.65
✓	✓		✓	95.93	81.11	88.27

Table 6: ACC (%) comparison of real images across models on training and test domains.

Method	FF++	Celeb-DF	DFDC	DFDCp	UADFV
Xception	85.14	27.61	61.76	44.68	55.10
EfficientNet	85.28	31.43	52.71	60.17	61.22
CADDM	87.93	26.29	60.62	70.49	81.63
Effort	88.50	37.11	60.63	71.59	85.71
RDBC	**90.68**	**39.62**	**62.82**	**80.89**	**91.84**

Table 7: Channel-response drift from CLIP. Lower values indicate better compatibility.

Model	**Head-30% W_2**	**Tail-10% W_2**	**Tail-30% W_2**	**Tail-30% Cov. drift**
CLIP	0.0000	0.0000	0.0000	0.0000
Effort	0.2879	0.2133	0.2121	0.5317
RDBC	**0.2734**	**0.2124**	**0.2035**	**0.5115**

4.2 Ablation Studies

Robustness against Image Degradations. Images always undergo various post-processing operations. To evaluate the robustness, we performed experiments on the Celeb-DF dataset with three common degradations: JPEG compression (Quality = 70 and 50), Gaussian blur ($\sigma = 3$), and Gaussian noise ($\sigma^2 = 0.01$). The results are presented in Table 4. As observed, baseline methods suffer performance drops when facing degradations. In contrast, our RDBC exhibits remarkable resilience across all types of corruptions, maintaining an AUC of 88.75% even under severe JPEG compression ($Q = 50$). This robustness may stem from our core motivation: RDBC preserves the statistical properties of real images encoded by the pretrained model. Since the pretrained statistical distribution of real data is inherently more stable and less sensitive to local pixel-level perturbations, our method consistently maintains the distinctiveness regardless of the applied physical degradations.

Module Analysis. We report the ablation study of the modules in RDBC in Table 5. RCE* uses the fitted Gaussian center (μ_0, C_0) during both training and inference, whereas RCE samples Gaussian-extrapolated statistics during training and uses the same center at inference. RCE* improves upon the backbone because the covariance-guided transformation allows multiple statistically related channels to participate in the representation, thereby reducing excessive reliance on isolated channel responses. Gaussian-fitted extrapolation (*i.e.*, *backbone + RCE*) further improves the DFDC AUC from 78.35% to 80.27%, supporting the benefit of training with continuous source-real-centered statistical references beyond a single fixed estimate.

Adding RKA alone improves AUC by an average of 1.00 percentage point across the three target domains. During fine-tuning, the frozen post-LN applies fixed channel-wise affine parameters to changing input features, while classification supervision primarily favors head channels. RKA uses the input-side kurtosis of each real sample to constrain its post-LN output, providing an additional statistical constraint on the representation entering the classifier. The gains obtained without RCE show that calibrating this final channel-response structure is beneficial on its own. Combining RCE with RKA achieves the best overall performance, showing that shallow covariance-guided channel enhancement and deep kurtosis-guided channel alignment provide complementary benefits.

Performance on Real Images. We report the performance of the models on real images in Table 6. Since this evaluation involves a single class, we use ACC as the primary metric. As shown, the baselines perform poorly on real images. For instance, Xception achieves only 27.61% accuracy on Celeb-DF. In contrast, our method is designed to model the distribution of real images, leading to significantly improved accuracy on real samples. On DFDCp, our approach improves real-image accuracy by 9.30 percentage points over the baseline (Effort).

Impact of Sample Size. The subset size n controls the reliability of the mean and log directional variance estimates used to fit the Gaussian models. We vary $n \in \{8, 16, 32, 64, 128\}$ to examine the sensitivity of RCE to this choice. As illustrated in Figure 3, small subsets (*e.g.*, $n = 8$) produce noisier statistical estimates and slightly lower cross-domain performance. The average cross-domain

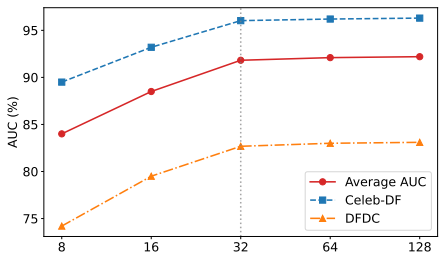


Figure 3: Effect of the subset size n used to fit the source-real statistical distributions.

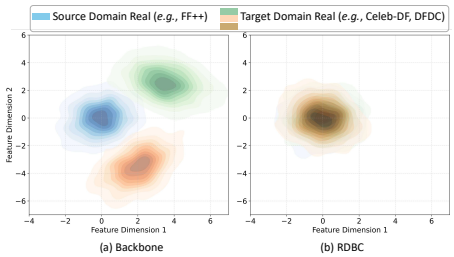


Figure 4: KDE contours show that different domain densities overlap closely under RDBC.

AUC stabilizes at $n = 32$, while increasing n to 64 or 128 yields only marginal gains. We use $n = 128$ in subsequent experiments to obtain more stable statistical estimates.

4.3 Empirical Analysis of Channel-Parameter Mismatch Correction

Post-LN Response Compatibility. We evaluate whether RDBC alleviates channel-parameter mismatch on the unseen domain. Since the affine parameters of post-LN remain unchanged, any nonzero response deviation results from the interaction between the shifted upstream features and the frozen pretrained mapping. We compute the empirical Wasserstein-2 distance and the relative covariance drift as provided in Appendix B. As shown in Table 7, RDBC reduces the response shift of the Head-30% channels from 0.2879 to 0.2734, the shifts of the Tail-10% and Tail-30% channels from 0.2133 and 0.2121 to 0.2124 and 0.2035, respectively, and the Tail-30% covariance drift from 0.5317 to 0.5115. These target-domain improvements demonstrate that our calibration transfers to unseen real images and maintains better statistical compatibility between channel responses and the frozen pretrained mapping. Note that these unnormalized Wasserstein-2 distances are affected by channel-specific scaling in the frozen affine transformation; their absolute magnitudes therefore do not directly indicate whether head or tail channels exhibit more severe relative mismatch.

Channel-level comparison. Figure 1 compares Effort and RDBC on the same images. In Figure 1(a), the top 0.1% of logit-contribution channels recover 99.42% source-domain AUC for Effort, compared with 95.08% for RDBC. This indicates that Effort relies more on a small group of dominant channels for discrimination. Figure 1(b) measures feature-gradient shares on the same channels as in Figure 1(a), and RDBC shows a lower gradient share in the head and a higher share in the tail than Effort. On target images, the lowest-contribution channels also behave differently: the lowest 30% have a mean signed contribution of -0.078 for Effort but $+0.075$ for RDBC (Figure 1(c)). These observations are consistent with RCE reducing reliance on isolated shallow responses and RKA calibrating the final channel structure.

Cross-domain real-feature visualization. Figure 4 compares two-dimensional KDE contours of real images from different domains. Under the backbone, the plotted densities occupy visibly different regions (Figure 4(a)). Under RDBC, their contours overlap more closely (Figure 4(b)). This qualitative pattern is consistent with our real-statistics-guided calibration, which reduces interference with target-domain classification caused by mismatches in tail-channel responses. These observations also suggest that the broad real-image knowledge encoded by pretrained models is worth exploiting for deepfake detection Wang et al. (2026b); Mu et al. (2026), while source-domain fine-tuning can shift the corresponding representations away from their pretrained statistical structure.

5 Discussion and Conclusion

Our analysis reveals a limitation of parameter-preserving fine-tuning: frozen weights do not preserve channel responses, while classification supervision favors a few head channels and leaves tail channels weakly constrained. These tail responses can interfere with predictions in unseen domains. Motivated by this observation, we propose RDBC, which comprises two complementary modules to use real-image statistics to calibrate shallow response concentration and accumulated channel-parameter mismatch at the classifier input. Experiments show that the two corrections complement each other, improving in-domain detection and cross-domain generalization.

AI USE STATEMENT

Generative AI was used to help refine the research framing and develop candidate mathematical objectives and explanatory arguments, design the evaluation protocol, search and summarize related literature, rewrite the English manuscript, and edit and compile its LaTeX source. The authors verified the method specification, arguments, references, and final empirical claims and are responsible for the final submitted content.

REFERENCES

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

Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. Voice-face homogeneity tells deepfake. *ACM ToMM*, 20(3):1–22, 2023.

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

Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *CVPR*, pp. 1133–1143, 2024.

Feng Ding, Wenhui Yi, Yunpeng Zhou, Xinan He, Hong Rao, and Shu Hu. Decoupling bias, aligning distributions: Synergistic fairness optimization for deepfake detection. In *CVPR*, pp. 32704–32713, 2026.

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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, pp. 1–13, 2024.

Hongyan Fei, Zexi Jia, Chuanwei Huang, Jinchao Zhang, and Jie Zhou. Exploring specular reflection inconsistency for generalizable face forgery detection. In *ICLR*, 2026.

Yue-Hua Han, Tai-Ming Huang, Kai-Lung Hua, and Jun-Cheng Chen. Towards more general video-based deepfake detection through facial component guided adaptation for foundation model. In *CVPR*, pp. 22995–23005, 2025.

Yikun Ji, Yan Hong, Qi Fan, Huijia Zhu, Weiqiang Wang, Liqing Zhang, Jianfu Zhang, et al. Fakexplain: Ai-generated image detection via human-aligned grounded reasoning. In *ICLR*, pp. 68549–68570, 2026.

Inho Jung, Hyeongjun Choi, Minh Binh Le, Hohyun Na, and Simon Woo. A rich knowledge space for scalable deepfake detection. In *ICLR*, pp. 96506–96535, 2026.

Shashank Kotyan, Makoto Shing, Yuki Imajuku, Rujikorn Charakorn, and Tarin Clanuwat. Prior-conditioned gaussian discriminants for generalizable ai-generated image detection. *arXiv preprint arXiv:2608.18523*, 2026.

Gert Lek, Nicolas Van Schaik, Chaoyi Zhu, Pin-Yu Chen, Robert Birke, and Lydia Chen. Detective sam: Adaptive ai-image forgery localization. In *ICLR*, pp. 22463–22490, 2026.

Hanzhe Li, Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, and Junyu Dong. Freqblender: Enhancing deepfake detection by blending frequency knowledge. In *NeurIPS*, volume 37, pp. 44965–44988, 2024.

Xiaotong Li, Yongxing Dai, Yixiao Ge, Jun Liu, Ying Shan, and Ling-Yu Duan. Uncertainty modeling for out-of-distribution generalization. *arXiv preprint arXiv:2202.03958*, 2022.

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

Zhengcen Li, Chenyang Jiang, Hang Zhao, Shiyang Zhou, Yunyang Mo, Feng Gao, Fan Yang, Qiben Shan, Shaocong Wu, and Jingyong Su. Preserving forgery artifacts: Ai-generated video detection at native scale. *arXiv preprint arXiv:2604.04634*, 2026.

Long Ma, Zhiyuan Yan, Jin Xu, Yize Chen, Qinglang Guo, Zhen Bi, Yong Liao, and Hui Lin. From specificity to generality: Revisiting generalizable artifacts in detecting face deepfakes. In *NeurIPS*, 2025.

Lianrui Mu, Haoji Hu, Xingze Zou, Jianhong Bai, and Jiaqi Hu. No pixel left behind: A detail-preserving architecture for robust high-resolution ai-generated image detection. In *ICLR*, pp. 80332–80355, 2026.

Saralees Nadarajah. Gaussian dct coefficient models. *Acta Appl. Math.*, 106(3):455–472, 2009.

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

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

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

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

Abu Taib Mohammed Shahjahan, Mohammad Mannan, Abdessamad Ben Hamza, and Amr Youssef. On improving robustness of deepfake image detectors. *arXiv preprint arXiv:2606.02797*, 2026.

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

Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annu. Rev. Neurosci.*, 24(1):1193–1216, 2001.

Ke Sun, Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, and Rongrong Ji. Diffusionfake: Enhancing generalization in deepfake detection via guided stable diffusion. In *NeurIPS*, pp. 101474–101497, 2024.

Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Towards general visual-linguistic face forgery detection. In *CVPR*, pp. 19576–19586, 2025.

Yuxia Sun, Ran Zhang, Aoxiang Sun, Xu Li, Zitao Liu, and Jingcai Guo. D&r: Recovery-based ai-generated text detection via a single black-box llm call. In *ICLR*, 2026.

Chuangchuang Tan, Xiang Ming, Jinglu Wang, Renshuai Tao, Bin Li, Yunchao Wei, Yao Zhao, and Yan Lu. Semantic visual anomaly detection and reasoning in ai-generated images. In *ICLR*, volume 2026, pp. 111907–111942, 2026a.

Hao Tan, Zichang Tan, Senyuan Shi, Ajian Liu, Chuanbiao Song, Huijia Zhu, Weiqiang Wang, Jun Wan, Zhen Lei, et al. Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *ICLR*, pp. 66420–66477, 2026b.

Jiahe Tian, Peng Chen, Cai Yu, Xiaomeng Fu, Xi Wang, Jiao Dai, and Jizhong Han. Learning to discover forgery cues for face forgery detection. *IEEE TIFS*, 19:3814–3828, 2024.

Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *CVPR*, pp. 14923–14932, 2021.

Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, and Yinglong Wang. Deepfake detection: A comprehensive survey from the reliability perspective. *ACM CSUR*, 57(3):1–35, 2024.

Yabin Wang, Zhiwu Huang, Zhou Su, Adam Prugel-Bennett, and Xiaopeng Hong. Penny-wise and pound-foolish in ai-generated image detection. *IEEE TPAMI*, pp. 1–14, 2026a.

Youqi Wang, Shen Chen, Haowei Wang, Rongxuan Peng, Taiping Yao, Shunquan Tan, Changsheng Chen, Bin Li, and Shouhong Ding. Forgeryvcr: Visual-centric reasoning via efficient forensic tools in mllms for image forgery detection and localization. *arXiv preprint arXiv:2602.14098*, 2026b.

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

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *CVPR*, pp. 8984–8994, 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 *ICML*, pp. 1–13, 2025a.

Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, Yunsheng Wu, and Li Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. In *CVPR*, pp. 12615–12625, 2025b.

Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP*, pp. 8261–8265, 2019.

Zheng Yang, Ruoxin Chen, Zhiyuan Yan, Ke-Yue Zhang, Xinghe Fu, Shuang Wu, Xiujun Shu, Taiping Yao, Shouhong Ding, Zequn Qin, et al. All patches matter, more patches better: Enhance ai-generated image detection via panoptic patch learning. In *ICLR*, pp. 121742–121761, 2026.

Zheyuan Zhang, Bin Wang, Debesh Jha, Ugur Demir, and Ulas Bagci. Domain generalization with correlated style uncertainty. In *WACV*, pp. 1989–1998, 2024.

Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

A GAUSSIAN-FITTED SOURCE-REAL STATISTICAL EXTRAPOLATION

This appendix details the statistical construction and theoretical properties of the Gaussian-fitted extrapolation used by RCE.

A.1 SUBSET STATISTICS AND GAUSSIAN FITTING

Let $\mathcal{I}_b \subseteq \{1, \ldots, N_r\}$ denote the b-th randomly drawn subset of n source-domain real features. Conditional on the collected source-real feature pool, each subset is sampled independently and uniformly; sampling is without replacement within a subset, whereas different subsets may overlap. Its mean and covariance estimates are

$$\mu_b = \frac{1}{n} \sum_{i \in \mathcal{I}_b} x_i^r, \quad C_b = \frac{1}{n-1} \sum_{i \in \mathcal{I}_b} (x_i^r - \mu_b)(x_i^r - \mu_b)^\top. \tag{15}$$

For i.i.d. features from a source-domain real distribution with finite fourth moments, the joint sampling distribution of the sample mean and the vectorized sample covariance is asymptotically Gaussian. Specifically, if $\mu_\star$ and $C_\star$ denote the statistics of this source-domain distribution, then

$$\sqrt{n} \begin{bmatrix} \mu_b - \mu_\star \\ \mathrm{vech}(C_b - C_\star) \end{bmatrix} \xrightarrow{d} \mathcal{N}(0, \Omega), \tag{16}$$

for an asymptotic covariance Ω. Here, $\mathrm{vech}(\cdot)$ stacks the lower-triangular entries of a symmetric matrix, including the diagonal. This population result motivates the Gaussian fitting in the main text. For subsets drawn without replacement from a finite feature pool, the sampling fraction also affects statistical variability; the i.i.d. approximation is most appropriate when n/N_r is small. We estimate the fitting variances directly from the B subset estimates. The fitted models describe local variation around source-domain real statistics, without assuming Gaussian real features or coverage of the entire real-image population.

We use the eigenspace of the full source-real covariance $C = Q \,\mathrm{diag}(\lambda_1, \ldots, \lambda_D) Q^\top$ to retain source-real channel dependencies. Conditioning on this fixed basis Q, the log directional variance for subset b and direction q_k is

$$s_{bk} = \log(q_k^\top C_b q_k + \epsilon), \qquad \epsilon > 0. \tag{17}$$

The mapping in Eq. (17) is smooth for $\epsilon > 0$; hence, the delta method motivates an asymptotic Gaussian approximation for each s_{bk}. From the B subset estimates, we calculate the maximum-likelihood parameters of factorized Gaussian distributions:

$$\begin{aligned} \bar{\mu}_j &= \frac{1}{B} \sum_{b=1}^{B} \mu_{b,j}, \qquad (\sigma_\mu)_j^2 = \frac{1}{B} \sum_{b=1}^{B} (\mu_{b,j} - \bar{\mu}_j)^2, \\ \bar{s}_k &= \frac{1}{B} \sum_{b=1}^{B} s_{bk}, \qquad (\sigma_s)_k^2 = \frac{1}{B} \sum_{b=1}^{B} (s_{bk} - \bar{s}_k)^2. \end{aligned} \tag{18}$$

The factorized form is a tractability approximation that avoids estimating a high-dimensional full covariance from only B subsets. The fixed basis Q preserves the principal directions of the source-real reference, while the sampled directional variances change their strengths. This construction uses the diagonal of $Q^\top C_b Q$ and does not retain its off-diagonal entries; the reconstructed covariance therefore shares the reference directions rather than reproducing each C_b exactly. We then sample

$$\widetilde{\mu}_j = \bar{\mu}_j + (\sigma_\mu)_j z_j^\mu, \quad \widetilde{s}_k = \bar{s}_k + (\sigma_s)_k z_k^C. \tag{19}$$

Here, z_j^μ and z_k^C are independent standard normal variables. Unlike empirical resampling from the B observed pairs, Eq. (19) defines a continuous distribution and can generate new combinations of source-real-centered statistics. This element-wise form is equivalent to the vector notation in Eq. (13).

A.2 Covariance validity and directional effect

The sampled covariance is reconstructed as

$$\widetilde{C} = Q \operatorname{diag}\left(e^{\widetilde{s}_1}, \ldots, e^{\widetilde{s}_D}\right) Q^\top. \tag{20}$$

For any nonzero $v \in \mathbb{R}^D$,

$$v^\top \widetilde{C} v = \sum_{k=1}^{D} e^{\widetilde{s}_k} (q_k^\top v)^2 > 0, \tag{21}$$

so every sampled $\widetilde{C}$ is symmetric positive definite. Moreover, its channel-wise entries satisfy

$$\widetilde{C}_{jl} = \sum_{k=1}^{D} q_{jk} q_{lk} e^{\widetilde{s}_k}, \tag{22}$$

which shows that varying the strength of a source-real covariance direction continues to couple its co-varying channels in the original channel coordinates.

Finally, condition on one sampled pair $(\widetilde{\mu}, \widetilde{C})$ and consider the shared RCE transformation $T(x) = x + \widetilde{C}(x - \widetilde{\mu})$ applied to the selected real and fake patch features. Let $d = \mu_f - \mu_r$ and $\hat{d} = \hat{\mu}_f - \hat{\mu}_r$ denote their mean differences before and after transformation. The common centering term cancels, yielding $\hat{d} = (I + \widetilde{C})d$. Therefore,

$$\|\hat{d}\|_2^2 - \|d\|_2^2 = 2d^\top \widetilde{C} d + \|\widetilde{C} d\|_2^2 \geq 0. \tag{23}$$

The inequality is strict whenever $\widetilde{C} d \neq 0$, which holds for any nonzero d when $\widetilde{C}$ is positive definite. The same derivation applies to the fixed positive-semidefinite C in Eq. (12), with equality possible when $Cd = 0$. Thus, at the selected-patch transformation stage, RCE amplifies components of an existing real-fake mean difference along source-real covariance directions. This property concerns transformation-stage mean separation and does not by itself guarantee a larger final classifier margin.

A.3 Deterministic inference

Stochastic extrapolation is used only during training. At inference, we use the centers of the fitted Gaussian distributions in the latent (μ, s) coordinate space:

$$\mu_0 = \bar{\mu}, \qquad C_0 = Q \operatorname{diag}\left(e^{\bar{s}_1}, \ldots, e^{\bar{s}_D}\right) Q^\top. \tag{24}$$

The same fixed pair (μ_0, C_0) is used for all test samples, removing test-time sampling variance from the reported evaluation. Here, C_0 is obtained by exponentiating the fitted log-variance centers; it is not the arithmetic mean of the sampled covariance matrices.

B Post-LN Channel-Parameter Compatibility Metrics

This appendix details the metrics used in Table 7. We process the same $N = 2750$ real images from the unseen DiffSwap domain with the original CLIP encoder, Effort, and RDBC. All models receive identical face crops and CLIP normalization. Let $x_i^m \in \mathbb{R}^D$, with $D = 1024$, denote the post-LN response of image i from model $m \in \{\mathrm{CLIP}, \mathrm{Effort}, \mathrm{RDBC}\}$. The original CLIP response serves as the reference. The post-LN affine transformation is

$$x_{ij}^m = \gamma_j u_{ij}^m + \beta_j, \qquad u_{ij}^m = \left[\mathrm{Norm}(x_i^{m,\mathrm{pre}})\right]_j, \tag{25}$$

where the same pretrained γ_j and β_j are frozen in all three models. We verify this directly by obtaining $\|\gamma^m - \gamma^{\mathrm{CLIP}}\|_\infty = 0$ and $\|\beta^m - \beta^{\mathrm{CLIP}}\|_\infty = 0$. Consequently, the measured post-LN response drift reflects how the changed upstream representation interacts with the fixed affine mapping rather than a change in the affine parameters themselves.

B.1 Fixed head and tail channel subsets

For a consistent comparison, we define one channel ordering using Effort and apply the same channel indices to all three models. Because all evaluated images are real, their correct-class label is $y = 0$. Using E as shorthand for Effort, let W_0^{E} and W_1^{E} denote its two classifier weight vectors. We rank channel j by its mean absolute correct-class margin contribution,

$$a_j = \frac{1}{N} \sum_{i=1}^{N} \left| x_{ij}^{\mathrm{E}} \left(W_{0,j}^{\mathrm{E}} - W_{1,j}^{\mathrm{E}} \right) \right|. \tag{26}$$

The Head-30% subset contains the channels with the largest a_j, whereas the Tail-10% and Tail-30% subsets contain those with the smallest a_j. We use the ceiling of each percentage of D, resulting in 308 Head-30%, 103 Tail-10%, and 308 Tail-30% channels. This target-domain ranking is used only for post-hoc analysis and does not update model parameters or provide target-domain supervision.

B.2 Empirical response-distribution drift

For each channel, we compare its empirical response distribution with the corresponding original CLIP distribution using the one-dimensional Wasserstein-2 distance. Let $x_{(i)j}^m$ denote the i-th order statistic obtained by sorting the N responses of channel j under model m. Because the same images are evaluated by every model, the empirical distance is

$$d_j^m = \left[\frac{1}{N} \sum_{i=1}^{N} \left(x_{(i)j}^m - x_{(i)j}^{\mathrm{CLIP}} \right)^2 \right]^{1/2}. \tag{27}$$

For a channel subset $\mathcal{S}$, the reported response drift is

$$D_{\mathrm{W2}}^m(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} d_j^m. \tag{28}$$

The original CLIP row is zero by construction. Moreover, Eq. (25) gives

$$W_2\big(x_j^m, x_j^{\mathrm{CLIP}}\big) = |\gamma_j| W_2\big(u_j^m, u_j^{\mathrm{CLIP}}\big), \tag{29}$$

where $W_2(x_j^m, x_j^{\mathrm{CLIP}})$ denotes the distance between the empirical distributions of the corresponding channel responses. The equality follows because both distributions undergo the same one-dimensional affine mapping. The metric therefore measures normalized-response drift after it is scaled by the corresponding frozen channel parameter.

We additionally measure whether the joint structure among tail channels remains compatible with the original CLIP representation. For model m and channel subset $\mathcal{S}$, let

$$C_{\mathcal{S}}^m = \frac{1}{N-1} \sum_{i=1}^{N} \left(x_{i,\mathcal{S}}^m - \bar{x}_{\mathcal{S}}^m \right) \left(x_{i,\mathcal{S}}^m - \bar{x}_{\mathcal{S}}^m \right)^\top. \tag{30}$$

Here, $\bar{x}_{\mathcal{S}}^m = N^{-1} \sum_i x_{i,\mathcal{S}}^m$ is the mean response vector over the evaluated images. The relative covariance drift reported for the Tail-30% subset is

$$D_{\mathrm{cov}}^m(\mathcal{S}) = \frac{\| C_{\mathcal{S}}^m - C_{\mathcal{S}}^{\mathrm{CLIP}} \|_F}{\| C_{\mathcal{S}}^{\mathrm{CLIP}} \|_F + \epsilon}, \qquad \epsilon = 10^{-12}. \tag{31}$$

These evaluation metrics are distinct from the RKA objective: RKA aligns sample-wise cross-channel kurtosis, whereas Eqs. (27)–(31) compare cross-image channel distributions and inter-channel covariance with the original CLIP reference. They therefore provide an independent assessment of post-LN response compatibility rather than restating the training loss.

C Interpreting Channel-Parameter Mismatch and Statistical Calibration

Parameter-preserving fine-tuning seeks to retain the real-image knowledge encoded by a pretrained foundation model. For a frozen mapping to use this knowledge after fine-tuning, its input responses

must remain sufficiently compatible with the mapping learned during pretraining; the responses themselves need not remain unchanged. Trainable and frozen parameter components can both affect the same channel, so head and tail channels do not correspond to separate sets of trainable and frozen weights. We use *channel-parameter mismatch* to describe a loss of statistical compatibility between a changed channel response and its downstream frozen mapping. The following analysis relates this possible mismatch to the observed weakness of tail-channel supervision and to the statistical calibration provided by RDBC.

C.1 Response drift through frozen mappings

Eq. (3) describes how an upstream feature change affects a frozen mapping. For a differentiable mapping and a small perturbation, its local form is

$$\Delta x_l = J_{f_{\text{fix}}}(x_{l-1})\delta_{l-1} + o(\|\delta_{l-1}\|_2), \tag{32}$$

where $J_{f_{\text{fix}}}$ is the Jacobian of the frozen mapping with respect to its input. Whenever the leading term is nonzero, a sufficiently small perturbation in that direction changes the output despite fixed parameters. For the affine part of post-LN, this relationship is exact: $\Delta x_{ij}^{\text{post}} = \gamma_j \delta_{ij}$, where δ_{ij} is the normalized-response change defined in Section 3.2. Thus, a nonzero change in a channel with $\gamma_j \neq 0$ produces a nonzero output change.

These relations establish that parameter freezing alone does not ensure response preservation. Continued changes to upstream parameters can keep altering the input of a frozen mapping, and the resulting response changes can propagate downstream when they remain insufficiently calibrated. Preserving pretrained real-image knowledge therefore calls for controlling the real responses presented to frozen mappings, rather than merely fixing their parameters. This motivates our use of real-image statistics to guide the changing representation while allowing the trainable component to learn task-related forgery evidence. Both parameter components can influence the same output channels.

C.2 Classification supervision and tail-channel interference

For the binary linear classifier, let p_i denote the predicted probability of the fake class. The per-sample cross-entropy gradient at the classifier input is

$$\frac{\partial \mathcal{L}_{\text{CE},i}}{\partial x_{ij}^{\text{post}}} = (p_i - y_i)(W_{1,j} - W_{0,j}). \tag{33}$$

Eq. (33) shows that the direct cross-entropy gradient for channel j depends on the classifier weight difference $W_{1,j} - W_{0,j}$ and the prediction error $p_i - y_i$. Once a small set of channels supports accurate source-domain classification, this objective need not strongly constrain the remaining responses. Figure 1(a,b) shows that channels with low source-domain logit contributions also receive low direct cross-entropy feature gradients. Shared upstream parameters can nevertheless continue to change these weakly supervised responses.

For a tail-channel subset $\mathcal{T}$, its contribution to the response-induced margin change in Eq. (6) is

$$\Delta M_{i,\mathcal{T}} = \sum_{j \in \mathcal{T}} (W_{y_i,j} - W_{1-y_i,j})\gamma_j \delta_{ij}. \tag{34}$$

The sum reduces the correct-class margin when it is negative, and small adverse terms can accumulate across channels. Figure 1(a,b) shows that the same low-contribution tail channels receive little direct gradient in the source domain, while panel (c) shows their effect on target-domain predictions. Together with the response drift in Table 7, these findings point to weakly constrained tail responses as a source of target-domain interference. Unlike a typical shortcut used for source-domain classification, these channels contribute little to source-domain predictions and receive limited direct classification supervision. Their behavior thus motivates examining how channel-parameter mismatch remains insufficiently corrected in the tail.

Table 1 also motivates examining channel structure early in the network: compared with the evaluated EfficientNet features, shallow CLIP responses have substantially greater cross-channel kurtosis and normalized response concentration. Their pronounced head–tail imbalance motivates calibration at shallow layers, before these responses propagate through the subsequent frozen mappings.

C.3 Complementary constraints from RCE and RKA

RDBC addresses this imbalance through complementary constraints on shallow and deep representations. RCE first acts on shallow high-response patches, transferring their centered deviations along channel dependencies estimated from source-domain real features. This encourages statistically related channels to participate in the representation and can reduce reliance on isolated high responses. The transformation also creates paths through which the loss can influence related channels. For a selected patch, holding the sampled statistics and patch selection fixed during differentiation gives

$$\nabla_{x_{ip}}\mathcal{L} = (I + \widetilde{C})^{\top}\nabla_{\hat{x}_{ip}}\mathcal{L}. \tag{35}$$

The off-diagonal covariance terms allow a channel to receive gradient contributions associated with other channels. By linking the responses and optimization signals of co-varying channels, RCE provides a route toward broader channel participation. For the same contribution-ranked channels, Figure 1(a,b) shows reduced head-channel reliance and a higher relative feature-gradient share in the low-contribution tail under RDBC. RCE uses real statistics to guide this transformation, while its shared operation on real and fake patches permits trainable parameters to develop discriminative fake responses.

RKA complements the shallow transformation at the final frozen post-LN. For real samples, it uses the current post-LN input kurtosis as a reference for the output kurtosis. This sample-wise statistic depends on the complete channel vector through its cross-channel mean, variance, and fourth moment. RKA therefore adds a distribution-level constraint involving both head and tail responses, strengthening the supervision of tail responses that receive little direct classification signal. Its purpose is to maintain a suitable real-image response structure across the frozen normalization interface while retaining the changes needed for deepfake detection. Together, RCE promotes broader channel participation, while RKA constrains the full response structure reaching the classifier. The real-only reference supports the use of retained pretrained knowledge; fake representations remain free of this direct alignment term and are shaped by the classification objective.

C.4 Evidence for joint head and tail improvement

Table 7 shows lower response-distribution drift for both head and tail subsets under RDBC. The Head-30% Wasserstein-2 distance decreases from 0.2879 to 0.2734, while the Tail-10% and Tail-30% distances decrease from 0.2133 to 0.2124 and from 0.2121 to 0.2035, respectively. The Tail-30% relative covariance drift also decreases from 0.5317 to 0.5115. These improvements show that RDBC maintains better statistical compatibility between both head and tail responses and the frozen pretrained mapping on unseen real images, without target-domain loss supervision.

Figure 1 provides complementary evidence. Panel (a) shows lower AUC recovery from a very small head-channel subset under RDBC, indicating reduced concentration of source-domain discrimination in those channels. Panel (b) measures the cross-entropy feature-gradient shares of the same channels ranked by source-domain logit contribution: low-contribution tail channels receive little direct gradient, but a higher relative share under RDBC than under Effort. Panels (a)–(c) use each model's source-domain logit-contribution ranking to characterize the source-domain contribution, direct gradient, and target-domain behavior of corresponding channel subsets.

Panel (c) further shows that low-contribution channels provide more support for correct-class prediction under RDBC on target images. Together, the table and figure support the interpretation that RDBC improves head and tail response compatibility while reducing the adverse effect of weakly constrained channels on target-domain predictions.