WHAT CLIP SEES WHEN STRUCTURE BREAKS: PATCH-SHUFFLE RESPONSE AS A GENERALIZABLE CUE FOR AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

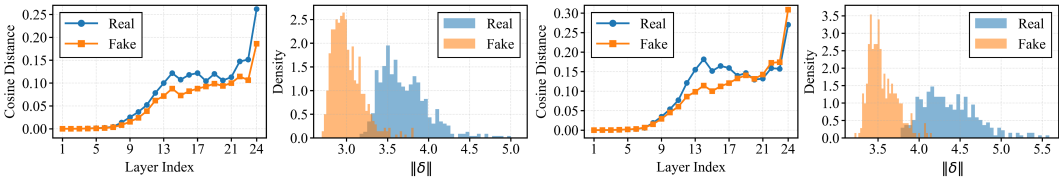
Existing CLIP-based detectors identify AI-generated images from the *static* representation of a single image, which is inevitably entangled with image semantics and thus generalizes poorly to unseen generators. In this work, we shift the focus from asking what an image *is* to how its representation *reacts* when its global structure is deliberately broken. Specifically, we compare an image's representation with that of its patch-shuffled counterpart, where global spatial arrangement is destroyed while local patches remain intact. We observe that real images exhibit significantly stronger response shifts than generated ones within a frozen CLIP feature space, suggesting that real representations rely heavily on globally coherent structures, whereas generated ones are dominated by local patterns. To harness this observation, we propose **SPARE** (**S**tructural **P**erturbation-**A**ware **RE**presentation), a framework that converts structural responses into a reliable discriminative signal, including a Patch-Shuffle structural response module constructs the perturbed view and computes layer-wise responses; a layer-wise response gating module adaptively weights each layer's contribution based on its reliability; and a dual-branch response fusion detector integrates the gated responses with the original representations as a content reference. Notably, the CLIP encoder remains entirely frozen throughout. Extensive experiments across multiple benchmarks demonstrate that SPARE consistently surpasses state-of-the-art detectors in cross-generator generalization.

1 INTRODUCTION

With the rapid advancement of deep generative models, particularly Generative Adversarial Networks (GANs) Goodfellow et al. (2020) and Diffusion Models Croitoru et al. (2023), synthesizing images with high visual fidelity and realistic details has become increasingly feasible. These generation techniques have enabled numerous applications in virtual reality, digital media, and entertainment. However, they have also introduced serious security concerns, including misinformation, identity forgery, and malicious content generation. Consequently, developing robust and generalizable detectors for AI-generated images has become crucial for safeguarding multimedia authenticity.

Early detection methods Marra et al. (2018); Yu et al. (2019); Frank et al. (2020); Wang et al. (2020a) primarily exploited low-level forensic artifacts, such as frequency anomalies, texture inconsistencies, and statistical irregularities. While effective for specific generators, these handcrafted cues often fail to generalize across different generation models. Recently, with the success of large-scale vision-language models Radford et al. (2021), CLIP-based detectors Ojha et al. (2023); Cozzolino et al. (2024); Liu et al. (2024a) have attracted considerable attention: the rich representations learned from web-scale image-text pairs provide more transferable features and improved cross-generator generalization. Nevertheless, existing CLIP-based detectors fundamentally derive their evidence from the *static* representation of a single image. UnivFD Ojha et al. (2023) probes the final-layer embedding; LTD Yang et al. (2026) mines the transitions between adjacent layers of these static features; SemAnti Chu et al. (2025) applies patch shuffling as a training-time regularization to suppress semantic bias, yet at inference it still classifies the static representation of the input image. In all cases, the discriminative signal remains a function of *what the image looks like* to the encoder, and

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) ProGAN (b) SD v1.4

Figure 1: **Structural responses separate real from synthetic images.** Given a frozen CLIP visual encoder, we compare the class token of an image with that of its patch-shuffled counterpart, and define their difference as the structural response $\boldsymbol{\delta}$ (Eq. 1). **Left:** layer-wise cosine distance between the two views; real images respond substantially more strongly, with the largest gap concentrated in mid-level layers (13–17). **Right:** distributions of $\|\boldsymbol{\delta}\|_2$ over the selected layers, showing clear training-free separation between real and generated images. The behavior is consistent for both ProGAN (a) and SD v1.4 (b), supporting our hypothesis that real representations depend on globally coherent structure whereas generated ones are dominated by local content.

such static representations are inevitably entangled with image semantics, leaving these detectors vulnerable to semantic distribution shifts across unseen domains and generators.

In this work, we propose a different perspective: rather than asking what an image *is*, we ask how its representation *reacts* when the global structure of the image is deliberately broken. Our key intuition is grounded in how the two classes of images come into being. A real photograph depicts a physical scene governed by consistent geometry, illumination, and perspective, so the appearance of every region is mutually constrained by its neighbors and by its position in the overall layout; the encoder's global representation of a real image thus relies heavily on the precise spatial arrangement of its parts. Generated images, in contrast, are synthesized through spatially local operations, and their global coherence arises as a soft statistical prior rather than a hard physical constraint. When a structural perturbation destroys the spatial arrangement while leaving local content intact, the representations of real and generated images therefore degrade in systematically different ways. Crucially, because the perturbation preserves the content within each patch, the shared content-related information is attenuated by the comparison between the original and perturbed representations. The resulting response reflects the representation's dependence on global structure rather than dataset-specific semantics, making it a naturally semantics-suppressed cue.

To verify this hypothesis, we conduct a preliminary analysis on a frozen CLIP visual encoder. We construct perturbed views via Patch Shuffle, an operation that randomly permutes image patches while keeping each patch intact, previously employed only as a training-time augmentation in CLIP-based detection Chu et al. (2025), and measure the representation discrepancy between the original and perturbed views as the structural response. As illustrated in Fig. 1, real and generated images exhibit clearly different response patterns: real images respond markedly more strongly, the separation is most pronounced in mid-level layers, and the same behavior holds across generators as different as ProGAN and SD v1.4, entirely without training. This observation confirms that structural responses capture the dependence of visual representations on global structural coherence, providing a discriminative cue that is markedly less affected by semantic content than static features.

Motivated by this observation, we propose **SPARE** (**S**tructural **P**erturbation-**A**ware **RE**presentation), a lightweight detection framework that elevates the structural response from an incidental byproduct to an explicit, first-class discriminative signal. Specifically, we first introduce a Patch-Shuffle Structural Response (PSR) module, which applies Patch Shuffle to construct the structural perturbation and measures the layer-wise representation differences between the original and perturbed views as structural responses. We then develop a Layer-wise Response Gating (LRG) module that adaptively estimates the contribution of the response at each selected Transformer layer and suppresses perturbation-induced noise. Finally, a Dual-branch Response Fusion (DRF) detector jointly encodes the original representations and the gated structural responses, where the former serves as a content-conditioned reference that allows the detector to judge whether the observed response is anomalous given the image content. The CLIP encoder remains entirely frozen throughout, and only several lightweight modules are trained.

Our main contributions are summarized as follows: (1) We introduce a new perspective for AI-generated image detection that shifts the discriminative evidence from static image representations

to *representation responses under structural perturbations*, and provide training-free empirical evidence that such responses separate real from generated images consistently across generators; (2) We propose SPARE, which realizes this perspective through three dedicated components: PSR for constructing and measuring structural responses, LRG for reliability-aware layer-wise weighting, and DRF for fusing responses with content-conditioned references, all built upon a frozen CLIP encoder with only several lightweight trainable modules; (3) Extensive experiments on multiple benchmarks demonstrate that SPARE achieves superior cross-generator generalization over state-of-the-art detectors, while requiring no backbone fine-tuning.

2 Related Works

Artifact-Driven Detectors. Early AI-generated image detectors mainly rely on identifying generation artifacts introduced by GAN-based models. Models such as ProGAN Karras et al. (2018) may introduce artifacts in frequency distributions, upsampling operations, and color consistency during synthesis. Therefore, CNN-based approaches have been developed to capture low-level discrepancies between real and generated images. For example, CNNSpot Wang et al. (2020a) employs ResNet to learn discriminative features associated with GAN-generated images. LGrad Tan et al. (2023) further exploits gradient-based representations from pretrained classifiers to highlight local abnormal responses and texture inconsistencies. However, with the rapid development of diffusion models, generated images become increasingly realistic with fewer explicit artifacts, limiting the generalization ability of artifact-driven detectors to advanced and unseen generators.

Local Forensic Cues Beyond Semantics. To overcome the limitations of artifact-driven detection, recent studies have explored more generalizable forensic cues beyond explicit artifacts and semantic information. PatchCraft Zhong et al. (2023) constructs texture-aware views through patch reorganization and exploits pixel correlation differences for detection. SemAnti Chu et al. (2025) adopts patch-shuffled training and selective fine-tuning to suppress semantic shortcuts while preserving CLIP knowledge. Zhou et al. Zhou et al. (2026) employ pixel-wise transformations to disrupt semantic structures and emphasize transferable generation traces. Different from these approaches that mainly rely on handcrafted perturbations or training strategies, we investigate CLIP representation responses under structural perturbations and explicitly model the response discrepancy between real and synthetic images for robust detection.

3 Methodology

3.1 Overview

We propose **SPARE** (**S**tructural **P**erturbation-**A**ware **RE**presentation), a lightweight framework that detects AI-generated images by measuring how a frozen visual encoder *responds* to a targeted structural perturbation, rather than by classifying its static representations alone. As shown in Figure 2, SPARE consists of three components. The Patch-Shuffle Structural Response (PSR) module constructs a perturbed view that destroys the global spatial arrangement while preserving local patch content, and defines the layer-wise representational discrepancy between the two views as the structural response. The Layer-wise Response Gating (LRG) module adaptively re-weights the response of each selected layer according to its reliability. The Dual-branch Response Fusion (DRF) detector jointly encodes the original representation and the gated structural response, and predicts image authenticity from their fused representation. The entire visual encoder remains frozen; only several lightweight modules are trained.

3.2 Preliminaries

CLIP-ViT Block Tokens. Given an input image $x \in \mathbb{R}^{H\times W\times C}$, we employ a frozen CLIP-ViT visual encoder $E(\cdot)$ to extract hierarchical visual representations. Let L denote the number of Transformer blocks. A forward pass through the encoder yields the intermediate representations from all blocks as $\{\mathbf{F}^1, \mathbf{F}^2, \ldots, \mathbf{F}^L\}$, where the representation at the ℓ-th block is denoted by $\mathbf{F}^\ell = [\mathbf{F}^\ell_{\text{cls}}, \mathbf{F}^\ell_{\text{patch}}]$. Here, $\mathbf{F}^\ell_{\text{cls}} \in \mathbb{R}^d$ denotes the class token embedding, and $\mathbf{F}^\ell_{\text{patch}} \in \mathbb{R}^{n\times d}$ represents the embeddings of n image patches, with d being the embedding dimension. Throughout this work, we build our structural response upon the class token. Because the class token aggregates

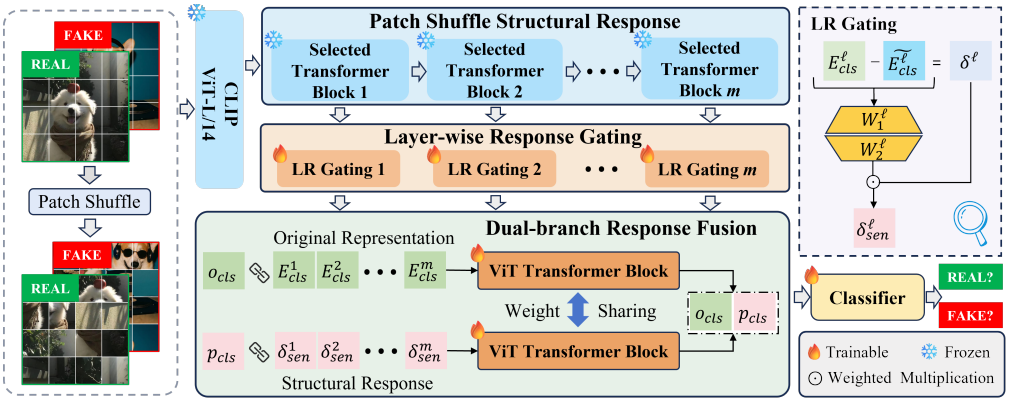


Figure 2: Overview of the proposed SPARE, consisting of three key components: Patch-Shuffle Structural Response (PSR) module for constructing structural responses, Layer-wise Response Gating (LRG) module for aggregating reliable responses, and Dual-branch Response Fusion (DRF) detector for final prediction.

information across all patches via self-attention, it directly reflects how the encoder integrates spatially distributed cues into a global representation, thereby serving as a highly sensitive indicator of shifts in global spatial configuration.

Layer Selection. Different Transformer blocks encode visual features at varying semantic levels, thus exhibiting distinct discriminative utility for AI-generated image detection. Following LTD Yang et al. (2026), we adopt the Adaptive Layer-wise Selection (ALS) module to automatically identify the most informative window of consecutive blocks. Formally, the selected layer index set is defined as $\mathcal{S} = \{s, s+1, \ldots, s+m-1\}$, where s denotes the starting layer index predicted by ALS and m is the window size (*i.e.*, the number of selected layers). Notably, $\mathcal{S}$ is derived solely from the original image and subsequently *reused* for the perturbed view to maintain structural consistency.

3.3 Patch Shuffle Structural Response

Motivation. Existing CLIP-based detectors search for discriminative evidence within the *static* representation of a single image. Such representations are inevitably entangled with semantic content, which drifts across datasets and generators and thus limits cross-domain generalization. We instead ask: rather than asking *what an image is*, we ask *how its representation reacts when its global structure is broken*.

We hypothesize that real and generated images differ systematically in this reaction. A real photograph depicts a physical scene governed by consistent geometry, illumination, and perspective; the appearance of any region is constrained by its neighbors and by its position in the overall layout. Consequently, the encoder's global representation of a real image relies heavily on the precise spatial arrangement of its parts. In contrast, generated images are synthesized through spatially local operations, and their global coherence arises as a soft statistical prior rather than a hard physical constraint. The dependence of their representations on the exact global arrangement is therefore comparatively weaker. When the arrangement is destroyed while local content is left intact, the resulting representational degradation is expected to follow different patterns for the two classes.

This formulation yields a desirable property for generalization. Because the perturbation preserves the content within each patch, the content-driven component that is common to both views is largely cancelled when the two representations are compared, leaving a signal that reflects arrangement dependence rather than dataset-specific semantics. Our empirical analysis in Fig. 1 supports this hypothesis: the magnitude of the structural response exhibits clearly separable distributions between real and generated images, and this separation remains stable across unseen generators.

Patch Shuffle Perturbation. To realize such a targeted perturbation, we employ Patch Shuffle. Given an input image x, Patch Shuffle first partitions the image into non-overlapping patches of equal size. A subset of patches is then randomly selected and their spatial positions are randomly permuted according to a predefined perturbation ratio. Denoting the Patch Shuffle operation as

PS$(\cdot)$, the perturbed image is formulated as $\tilde{x} = \text{PS}(x, \rho)$ with $\rho \in (0, 1]$, where ρ controls the proportion of shuffled patches. Patch Shuffle alters only the spatial arrangement without modifying the visual content within any patch. The perturbed image $\tilde{x}$ therefore preserves low-level visual information such as local textures, colors, and generator-specific traces, while disrupting the global spatial layout and long-range contextual dependencies. This selectivity is what makes Patch Shuffle a *targeted* rather than an arbitrary perturbation.

Structural Response. The original image x and its perturbed counterpart $\tilde{x}$ are fed into the same frozen CLIP-ViT visual encoder. Let $\mathbf{F}_{\text{cls}}^{\ell_j}$ and $\tilde{\mathbf{F}}_{\text{cls}}^{\ell_j}$ denote the class token representations extracted from the j-th selected Transformer layer for x and $\tilde{x}$, respectively. We define the *structural response* at this layer as

$$\boldsymbol{\delta}^j = \mathbf{F}_{\text{cls}}^{\ell_j} - \tilde{\mathbf{F}}_{\text{cls}}^{\ell_j}, \qquad j \in \{1, \ldots, m\}, \tag{1}$$

where $\boldsymbol{\delta}^j \in \mathbb{R}^d$. Following the reasoning above, $\boldsymbol{\delta}^j$ characterizes the extent to which the encoder's global representation at this depth depends on the spatial arrangement of the image, and thus serves as our basic discriminative unit.

3.4 Layer-wise Response Gating

Different Transformer blocks encode structural information with varying discriminative capability, and the structural responses from different layers contribute unequally. Moreover, Patch Shuffle inevitably induces responses that are unrelated to authenticity, which manifest as noise in certain layers. It is thus desirable to adaptively weight the structural responses according to their reliability.

To this end, we equip each position of the selected window with a lightweight gating network $G^j(\cdot)$ that estimates the contribution weight of the corresponding structural response. Note that the gating networks are bound to the *relative position* j within the selected window rather than to absolute layer indices, so that they remain valid under the dynamic selection produced by ALS. For each position, we first compute a compact statistical descriptor of the original and perturbed class tokens:

$$\mathbf{c}^j = \left[\mu(\mathbf{F}_{\text{cls}}^{\ell_j}), \|\mathbf{F}_{\text{cls}}^{\ell_j}\|_2, \mu(\tilde{\mathbf{F}}_{\text{cls}}^{\ell_j}), \|\tilde{\mathbf{F}}_{\text{cls}}^{\ell_j}\|_2\right] \in \mathbb{R}^4, \tag{2}$$

where $\mu(\cdot)$ denotes the mean over the feature dimension and $\|\cdot\|_2$ the L_2 norm. We deliberately derive the descriptor from the two views themselves rather than from $\boldsymbol{\delta}^j$: these statistics summarize the activation level of each view and the representational shift induced by the perturbation, while remaining complementary to the response being weighted.

The descriptor is then fed into the corresponding gating network to produce a scalar weight as $w^j = G^j(\mathbf{c}^j) \in [0, 1]$, where each G^j consists of two linear layers with a GELU activation followed by a Sigmoid function. Finally, the gated structural responses of all selected positions are stacked in layer order to form the structural response sequence:

$$\mathbf{P}^{\text{str}} = \text{Stack}\left(w^j \boldsymbol{\delta}^j \;\middle|\; j = 1, \ldots, m\right) \in \mathbb{R}^{m \times d}, \tag{3}$$

where $\text{Stack}(\cdot)$ denotes stacking along layer dimension. Through this gating mechanism, the model adaptively adjusts the contribution of each layer's structural response, suppressing unreliable responses and perturbation-induced noise, producing stable structural representations.

3.5 Dual-branch Response Fusion

Why two branches. The structural response cannot be interpreted in isolation. Its magnitude and direction are inevitably modulated by image content: a texture-rich, structurally complex image naturally exhibits a larger response than a smooth one, irrespective of its authenticity. Judging $\boldsymbol{\delta}$ alone therefore conflates content-driven variation with authenticity-driven variation. We address this by introducing a second branch that encodes the original representation, which serves not as the primary discriminative pathway but as a *content-conditioned reference frame*: it allows the detector to assess whether the observed structural response is anomalous *given* the content of the image. The two branches thus characterize the input from complementary perspectives.

Branch construction. We stack the class tokens of the selected layers in hierarchical order to obtain the original representation sequence:

$$\mathbf{P}^{\text{ori}} = \text{Stack}\left(\mathbf{F}_{\text{cls}}^{\ell_j} \;\middle|\; j = 1, \ldots, m\right) \in \mathbb{R}^{m \times d}. \tag{4}$$

Table 1: ACC (%) comparison on ForenSynths. Each row corresponds to a detection method, and each column to a generative model. The last column reports the mean ACC across all models. **Red** and blue indicate the best and second-best results, respectively.

Method	Generative Adversarial Networks							Deepfakes	SAN	LDM			Guided	Glide			Dalle	Mean ACC
	Pro-GAN	Style-GAN	Style2-GAN	Big-GAN	Cycle-GAN	Star-GAN	Gau-GAN			100	200	200_cfg		50_27	100_10	100_27		
CNNSpot$_{CVPR'20}$	93.00	62.00	52.97	57.00	69.61	77.94	62.99	59.41	51.83	52.60	52.90	52.20	51.80	55.20	54.10	54.30	50.25	59.42
UnivFD$_{CVPR'23}$	99.65	84.55	74.95	94.75	98.20	95.95	99.45	68.65	56.50	95.15	94.55	74.15	70.00	79.20	78.05	78.65	87.45	84.11
NPR$_{CVPR'24}$	99.10	87.90	89.00	79.40	71.50	60.60	66.60	71.00	69.90	73.20	73.20	78.00	69.00	92.00	93.00	91.20	61.40	78.00
FatFormer$_{CVPR'24}$	99.89	97.13	**98.30**	99.50	99.36	99.75	99.43	93.25	68.04	**98.65**	98.60	94.90	76.05	94.65	94.20	94.35	**98.75**	94.40
RINE$_{ECCV'24}$	99.90	84.10	76.90	98.00	99.40	99.60	99.70	69.90	63.00	98.40	98.00	86.40	64.00	84.80	83.20	82.20	93.70	87.13
FreqNet$_{AAAI'24}$	99.80	83.70	82.10	89.10	90.20	79.50	90.30	70.80	69.40	92.00	91.30	92.30	64.20	72.90	74.20	69.70	96.70	82.84
C2P-CLIP$_{AAAI'25}$	99.90	92.40	85.30	98.60	**99.90**	**100.00**	99.50	76.90	77.20	97.30	97.60	91.50	58.00	92.70	90.00	89.40	95.50	90.69
D$^3_{CVPR'25}$	96.62	63.84	54.99	82.80	88.04	66.33	95.19	51.95	47.73	53.07	52.40	37.67	39.27	46.90	45.93	46.37	45.50	59.68
VIB$_{CVPR'25}$	**100.00**	89.50	86.00	95.00	99.50	97.50	**100.00**	82.50	73.00	97.50	96.50	78.50	71.00	92.00	82.50	88.50	91.00	89.44
ForgeLens$_{ICCV'25}$	**100.00**	94.08	95.77	**99.67**	99.70	99.97	99.71	89.22	88.19	98.40	**98.80**	94.55	69.65	97.30	96.60	94.70	98.00	94.96
LTD$_{CVPR'26}$	99.86	94.42	88.78	99.05	99.47	99.97	99.30	**93.32**	84.93	93.40	93.10	85.90	49.30	96.40	94.30	93.40	97.20	91.89
SPARE (Ours)	99.96	**98.23**	95.35	91.25	94.66	99.80	90.67	91.14	**92.92**	97.20	97.40	**97.30**	**90.50**	**97.50**	**97.40**	**96.90**	97.95	**95.66**

A learnable class token is prepended to each branch input, and learnable positional embeddings are added to preserve the hierarchical ordering among the selected layers. The two branch inputs are formulated as

$$\mathbf{Z}^{\text{ori}} = \left[\mathbf{Z}^{\text{ori}}_{\text{cls}};\ \mathbf{P}^{\text{ori}} + \mathbf{Z}^{\text{ori}}_{\text{pos}}\right], \tag{5}$$

$$\mathbf{Z}^{\text{str}} = \left[\mathbf{Z}^{\text{str}}_{\text{cls}};\ \mathbf{P}^{\text{str}} + \mathbf{Z}^{\text{str}}_{\text{pos}}\right], \tag{6}$$

where $\mathbf{Z}^{\text{ori}}_{\text{cls}}$ and $\mathbf{Z}^{\text{str}}_{\text{cls}}$ are the learnable class tokens of the two branches, and $\mathbf{Z}^{\text{ori}}_{\text{pos}}, \mathbf{Z}^{\text{str}}_{\text{pos}} \in \mathbb{R}^{m\times d}$ are learnable positional embeddings indexed by the relative position within the selected window.

Both sequences are processed by a parameter-shared Transformer module $\mathcal{T}_\theta$:

$$\mathbf{H}^{\text{ori}} = \mathcal{T}_\theta(\mathbf{Z}^{\text{ori}}), \qquad \mathbf{H}^{\text{str}} = \mathcal{T}_\theta(\mathbf{Z}^{\text{str}}). \tag{7}$$

Sharing $\mathcal{T}_\theta$ across the two branches maps the original representations and the structural responses into a common space, which stabilizes optimization and prevents the two branches from drifting into incompatible feature geometries.

Prediction. The class token representations of the two branch outputs are concatenated and fed into a binary classification head:

$$\hat{y} = \sigma\left(\text{FC}\left([\mathbf{H}^{\text{ori}}_{\text{cls}};\ \mathbf{H}^{\text{str}}_{\text{cls}}]\right)\right), \tag{8}$$

where $\sigma(\cdot)$ is the Sigmoid function and $\text{FC}(\cdot)$ is the classification head.

3.6 Training Objective

The prediction is supervised by a binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{B}\sum_{b=1}^{B}\left[y_b \log \hat{y}_b + (1-y_b)\log(1-\hat{y}_b)\right], \tag{9}$$

where B is the batch size, $y_b \in \{0,1\}$ is the ground-truth label with 1 denoting a generated image, and $\hat{y}_b$ is the predicted probability for the b-th sample. SPARE requires no auxiliary objective, no adversarial training, and no synthetic hard-negative construction; the structural response is obtained directly from the frozen encoder and is supervised end-to-end by the detection loss alone.

4 Experiment

4.1 Experiment Settings

Datasets. Our proposed SPARE method is evaluated on three widely-used benchmarks, including **ForenSynths** Wang et al. (2020b), **GenImage**Zhu et al. (2023), and **DRCT-2M** Chen et al. (2024). Specifically, we adopt ForenSynths and DRCT-2M for training. For ForenSynths, we use a two-class subset (*tvmonitor* and *diningtable*) containing 72K images, including ProGAN-generated

Table 2: ACC (%) comparison on DRCT-2M. Each row corresponds to a detection method, and each column to a diffusion model variant. The last column reports the mean ACC across all models. **Red** and blue indicate the best and second-best results, respectively.

Method	SD Variants						Turbo Variants		LCM Variants		ControlNet Variants			Mean ACC
	LDM	SD v1.4	SD v1.5	SD v2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SD v1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	
CNNSpot$_{CVPR'20}$	80.80	82.53	82.14	75.99	52.82	53.86	54.03	54.12	61.90	52.64	62.08	53.82	66.97	64.13
UnivFD$_{CVPR'23}$	93.55	89.45	89.75	87.55	81.75	82.70	76.50	74.10	83.20	80.35	81.85	71.65	73.50	81.99
NPR$_{CVPR'24}$	67.20	**100.00**	**100.00**	57.50	50.40	60.90	**100.00**	**100.00**	**100.00**	56.20	56.40	53.20	56.20	73.69
FatFormer$_{CVPR'24}$	55.33	48.55	48.56	48.56	48.55	48.61	48.59	48.55	48.68	50.25	50.03	50.11	61.09	50.42
RINE$_{ECCV'24}$	94.50	96.50	96.50	95.80	84.40	71.20	75.20	75.60	89.80	77.90	90.40	66.30	83.50	84.43
FreqNet$_{AAAI'24}$	**97.40**	**100.00**	**100.00**	50.30	50.00	50.00	**100.00**	**100.00**	**100.00**	50.00	52.80	51.60	52.30	73.42
C2P-CLIP$_{AAAI'25}$	78.20	91.81	91.64	91.00	82.60	90.36	82.15	81.52	82.40	76.52	87.68	81.11	84.73	84.75
D^3 $_{CVPR'25}$	78.88	98.99	98.79	96.21	78.82	75.17	79.28	77.95	96.14	82.30	92.17	85.13	78.75	86.04
VIB$_{CVPR'25}$	89.50	59.50	60.00	64.00	65.50	56.50	61.00	57.00	57.00	70.00	77.50	74.50	78.00	66.92
LTD$_{CVPR'26}$	75.23	99.95	99.96	88.79	84.89	84.08	99.95	99.95	99.92	91.91	96.58	91.47	93.98	92.82
SPARE (Ours)	81.81	99.85	99.85	**97.43**	**97.91**	**97.10**	99.84	99.85	99.83	**98.97**	**98.92**	**96.54**	**98.34**	**97.40**

Table 3: ACC (%) performance comparison on GenImage. Each row corresponds to a detection method, and each column to a generative model. The last column reports the mean Acc across all models. **Red** and blue indicate the best and second-best results, respectively.

Method	ADM	Midjourney	GLID	VQDM	SDv1.4	SDv1.5	Wukong	BigGAN	Mean ACC
CNNSpot$_{CVPR'20}$	52.49	53.18	53.25	50.92	50.43	50.49	50.70	50.45	51.49
UnivFD$_{CVPR'23}$	66.70	56.15	62.00	85.45	63.85	63.35	70.70	90.25	69.81
NPR$_{CVPR'24}$	71.50	**87.20**	86.80	61.10	77.80	77.40	67.90	91.60	77.66
FatFormer$_{CVPR'24}$	78.47	56.08	88.05	86.89	67.83	68.04	73.07	96.83	76.91
RINE$_{ECCV'24}$	67.50	56.40	68.10	83.80	76.00	75.50	77.60	88.80	74.21
FreqNet$_{AAAI'24}$	81.30	62.90	62.10	75.30	52.40	52.90	50.70	81.60	64.90
C2P-CLIP$_{AAAI'25}$	59.86	55.04	79.58	75.64	80.85	80.86	78.30	95.43	75.70
D^3 $_{CVPR'25}$	57.28	50.04	55.78	69.63	50.63	50.58	51.58	80.25	58.22
VIB$_{CVPR'25}$	72.00	57.50	66.50	83.00	70.00	71.00	77.50	88.50	73.25
ForgeLens$_{ICCV'25}$	77.52	56.96	91.77	82.07	86.51	87.12	82.58	**99.39**	82.99
LTD$_{CVPR'26}$	82.75	59.58	95.45	87.91	92.04	92.37	91.42	96.92	87.31
SPARE (Ours)	**91.14**	70.23	**95.89**	**96.11**	**95.76**	**94.70**	**95.56**	96.33	**91.96**

images and real images collected from LSUN Yu et al. (2015). For DRCT-2M, we follow the official protocol and use 236K images consisting of SDv1.4-generated images and real images from MSCOCO Lin et al. (2014). For testing, ForenSynths provides synthetic images generated by 13 different GAN and DM generators. We further evaluate cross-dataset generalization by training on ForenSynths and testing on GenImage, which contains fake images generated by eight representative GAN and diffusion models, with real images collected from ImageNet Deng et al. (2009). Following the official protocol of DRCT-2M, we exclude the DR subset and use the remaining 13 synthetic subsets for testing.

Metrics and baseline methods. Following previous methods, we adopt Accuracy (ACC) and average precision (AP) metrics to evaluate the performance. The AP results are provided in Appendix A.1. For a fair comparison, the decision threshold is set to 0.5. We compare our SPARE with several representative baseline methods including CNNSpot Wang et al. (2020a), UnivFD Ojha et al. (2023), NPR Tan et al. (2024a), FatFormer Liu et al. (2024b), RINE Koutlis & Papadopoulos (2024), FreqNet Tan et al. (2024b), C2P-CLIP Tan et al. (2025), D^3 Yang et al. (2025), VIB Zhang et al. (2025), ForgeLens Chen et al. (2025) and LTD Yang et al. (2026).

Implementation details. Our method is implemented based on the PyTorch framework with a pre-trained CLIP-ViT/L-14 model as the backbone network. All experiments are conducted on an NVIDIA RTX 3090 GPU. During training, the parameters of the CLIP visual encoder are fully frozen, and only the lightweight detection module is optimized. For data preprocessing, all input images are first resized to 256×256 pixels, followed by random cropping to 224×224 pixels and random horizontal flipping for data augmentation. During inference, we apply center cropping to obtain images with 224×224 pixels. We train the model using the Adam optimizer with a learning rate of 5×10^{-5}, a batch size of 256, and 5 training epochs. The hyperparameter ρ is set to 0.7.

4.2 Comparison Against State-of-the-Art

Comparisons on ForenSynths. We report the ACC performance on the ForenSynths dataset across 17 generative models in Table 1. Overall, SPARE achieves the highest mean ACC of 95.66%, sur-

Table 4: Ablation study of the designed modules, reported in Mean ACC (%).

PSR	LRG	DRF	ForenSynths	GenImage	Average
×	×	×	86.54	68.44	77.49
✓	×	×	87.32	74.99	81.16
✓	✓	×	88.16	85.93	87.05
✓	×	✓	95.45	89.82	92.64
✓	✓	✓	**95.66**	**91.96**	**93.81**

Table 5: Ablation study of different perturbation strategies, reported in Mean ACC (%).

Perturbation	ForenSynths	GenImage	Average
Gaussian Noise	95.56	90.08	92.82
Patch Mask	88.25	86.77	87.51
Patch Shuffle	**95.66**	**91.96**	**93.81**

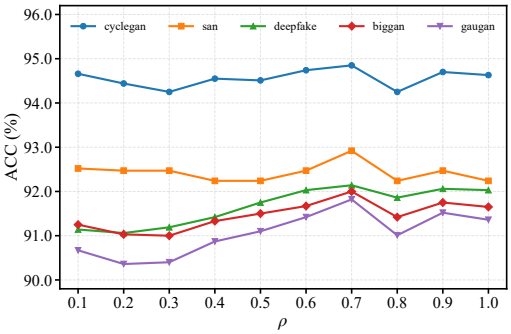


Figure 3: Effect of Patch Shuffle ratio ρ on five-class subset of ForenSynths.

passing the second-best method ForgeLens Chen et al. (2025) by 0.70%. This demonstrates that SPARE can effectively capture discriminative and generalizable synthesis cues across diverse image generators. Notably, SPARE maintains strong detection performance on both GAN-based generators, such as ProGAN and StyleGAN, and DM-based generators, such as LDM and Glide, indicating its robustness against different generation mechanisms. For challenging generators including SAN and Guided, SPARE achieves over 90% ACC, significantly outperforming existing methods. These results verify that SPARE learns generator-agnostic structural responses rather than relying on generator-specific artifacts, leading to improved generalization across unseen synthesis models.

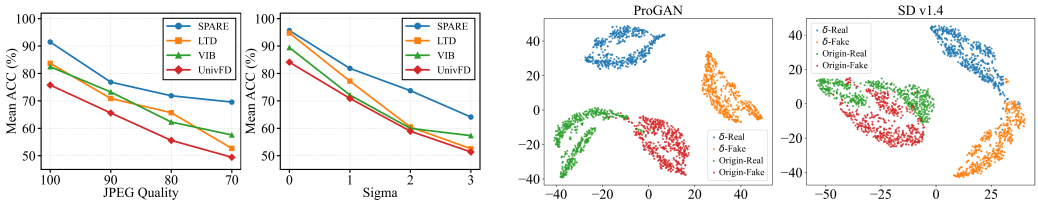
Comparisons on GenImage. Table 3 illustrates the cross-dataset ACC performance on GenImage. All the methods are trained on ForenSynths for fair comparison. As can be seen, SPARE achieves the best detection performance on most generative models. Specifically, SPARE achieves the highest mean ACC of 91.96%, surpassing the second-best method LTD Yang et al. (2026) by 4.66%. This result verifies that, even without exposure to the target test distribution during training, SPARE can effectively capture transferable synthesis patterns rather than generator-specific artifacts, leading to superior generalization.

Comparisons on DRCT-2M. We further evaluate the robustness of SPARE against diverse diffusion-based generation variants on DRCT-2M. As shown in Table 2, SPARE gains a mean ACC improvement of 4.58% compared to the best baseline LTD. Specifically, SPARE maintains excellent detection performance across diverse diffusion variants, including SDv1.4, SDv1.5, SDv2, SDXL with Refiner/Turbo extensions, and LCM- and ControlNet-based models, achieving nearly 100% accuracy on most generation settings. This demonstrates that SPARE can effectively capture shared synthesis patterns across different diffusion architectures and generation strategies, leading to strong cross-model generalization and robustness.

Robustness evaluation. In real-world scenarios, images are often affected by various degradations, such as compression artifacts and noise during transmission and storage. To evaluate the robustness of our method, we conduct experiments under different degradation levels, focusing on two common types of image perturbations: (1) JPEG compression, where the quality factor is gradually reduced from 100 to 70; and (2) Gaussian noise, where the noise intensity σ is increased from 0 to 3. As shown in Figure 4a, our proposed SPARE consistently achieves superior robustness compared with recent state-of-the-art methods. Specifically, SPARE maintains stable detection performance under various degradation conditions, exhibiting superior resilience to compression distortions and noise perturbations, which highlights its robustness in realistic scenarios.

4.3 Ablation Study

Effect of the designed modules. We analyze the impact of three components in SPARE: PSR introduces structural response for detection, LRG performs layer-wise response gating, and DRF adaptively integrates the base and the structure branches. The corresponding mean ACC results on ForenSynths and GenImage are reported in Table 4. Several observations can be drawn from the results: ***(1) PSR introduces discriminative structural response for detection.*** Introducing PSR improves the mean ACC by 3.67% over the baseline without using our proposed modules, indicating



(a) Robustness under JPEG and Gaussian noise.

(b) t-SNE visualization on ProGAN and SD v1.4.

Figure 4: Robustness evaluation and feature representation visualizations. (a) SPARE demonstrates strong robustness against realistic image degradations. (b) t-SNE visualization of original and response representations.

that structural response provides additional synthesis-related cues. ***(2) LRG effectively aggregates complementary responses across different layers.*** Compared with the baseline using PSR module, incorporating LRG further improves the detection performance, verifying its effectiveness in aggregating complementary responses from different layers. ***(3) DRF enables content-aware calibration.*** Compared with the baseline using PSR and LGR, incorporating DRF further improves the detection performance, demonstrating that the base branch provides content-aware references to calibrate structural responses and improve detection reliability.

Effect of perturbation strategies. To investigate the influence of different perturbation strategies, we compare the proposed patch shuffle with two alternative variants: (1) *Gaussian Noise*, which adds pixel-level Gaussian noise sampled from a standard normal distribution to the input image, and (2) *Patch Mask*, which randomly masks image patches by setting their pixel values to zero. As shown in Table 5, patch shuffle achieves the best performance among all perturbation strategies. This is because Gaussian noise mainly disturbs low-level appearance, while patch masking removes local information and introduces missing-content artifacts. In contrast, patch shuffle disrupts global spatial structures while preserving local semantics, providing more effective perturbation-induced responses.

Effect of patch shuffle ratio ρ. To investigate the effect of perturbation strength, we vary the Patch Shuffle ratio ρ from 0.1 to 1.0 with a step size of 0.1 and report the ACC on a five-class subset of ForenSynths in Figure 3. The performance remains relatively stable across a wide range of perturbation ratios, indicating that the proposed response modeling is robust to variations in perturbation strength. Moreover, most categories achieve their best at $\rho = 0.7$, which is therefore adopted as the default setting in our experiments.

4.4 T-SNE VISUALIZATION

To validate the effectiveness of the structural response representations, we visualize the feature distributions of the original and response branches. Figure 4b illustrates the t-SNE results of real and fake images sampled from ProGAN and SD v1.4. As observed, the original static features exhibit noticeable distribution overlap between real and fake samples, especially for SD v1.4, indicating that static representations may contain entangled distribution-specific information. In contrast, our structural response representations achieve clearer class boundaries and better separation between real and fake images. This demonstrates that the response representation alleviates the reliance on static feature biases and captures more generalizable forensic patterns.

5 CONCLUSION

This work reveals that synthetic and real images exhibit distinct sensitivities to structural perturbations, providing a reliable structural response-based cue for AI-generated image detection. Based on this perspective, we propose SPARE, which extracts reliable structural responses from patch-shuffled views and integrates them with original representations through adaptive response modeling for robust detection. Extensive experiments on multiple benchmarks demonstrate that SPARE achieves superior cross-generator generalization and consistently outperforms existing state-of-the-art detectors.

6 AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the manuscript and for retrieval and discovery of relevant literature.

REPRODUCIBILITY STATEMENT

Appendix A.1 provides AP results on ForenSynths, GenImage, and DRCT-2M datasets. Appendix A.2 analyzes the effect of patch shuffle randomness. Appendix A.3 reports parameter efficiency and inference speed. The code will be released with the final version.

REFERENCES

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024.

Yingjian Chen, Lei Zhang, and Yakun Niu. Forgelens: Data-efficient forgery focus for generalizable forgery image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16270–16280, 2025.

Beilin Chu, Weike You, Mengtao Li, Tingting Zheng, Kehan Zhao, Xuan Xu, Zhigao Lu, Jia Song, Moxuan Xu, and Linna Zhou. When semantics regulate: Rethinking patch shuffle and internal bias for generated image detection with CLIP. *arXiv preprint arXiv:2511.19126*, 2025.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4356–4366, 2024.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9): 10850–10869, 2023.

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

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3247–3258. PMLR, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 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.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on computer vision*, pp. 394–411. Springer, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

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, 2024a.

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

Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval*, pp. 384–389, 2018. doi: 10.1109/MIPR.2018.00084.

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.

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.

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, 2024a.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5052–5060, 2024b.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 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 *CVPR*, 2020a.

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, 2020b.

Yawen Yang, Feng Li, Shuqi Kong, Yunfeng Diao, Xinjian Gao, Zenglin Shi, and Meng Wang. Layer consistency matters: Elegant latent transition discrepancy for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. D3: Scaling up deepfake detection by learning from discrepancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Ning Yu, Larry S. Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7556–7566, 2019.

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 Computer Vision and Pattern Recognition Conference*, pp. 23828–23837, 2025.

Table 6: AP (%) comparison on ForenSynths. Each row corresponds to a detection method, and each column to a generative model. The last column reports the mean AP across all models. **Red** and blue indicate the best and second-best results, respectively.

Method	Generative Adversarial Networks							Deepfakes	SAN	LDM			Guided	Glide			Dalle	Mean AP
	Pro-GAN	Style-GAN	Style2-GAN	Big-GAN	Cycle-GAN	Star-GAN	Gau-GAN			100	200	200_cfg		50_27	100_10	100_27		
CNNSpot$_{CVPR'20}$	99.00	92.00	76.31	76.00	85.00	92.00	91.00	71.42	61.77	67.73	67.08	65.01	59.96	75.48	72.28	71.89	51.76	75.04
UnivFD$_{CVPR'23}$	**100.00**	97.38	97.86	99.17	99.76	99.38	99.98	81.82	78.62	99.35	99.40	93.22	88.27	96.03	95.48	95.79	97.73	95.25
NPR$_{CVPR'24}$	**100.00**	98.60	99.50	87.70	89.20	76.20	79.00	78.00	81.70	98.10	98.10	98.70	80.20	99.40	99.50	99.50	95.40	91.69
FatFormer$_{CVPR'24}$	**100.00**	99.75	**99.92**	**99.98**	**100.00**	**100.00**	**100.00**	97.99	81.23	99.87	99.81	99.09	92.01	99.41	99.21	99.13	99.82	98.07
RINE$_{ECCV'24}$	**100.00**	99.10	98.50	99.90	**100.00**	**100.00**	**100.00**	95.60	89.00	99.90	99.90	98.20	91.30	98.10	97.60	97.50	99.40	97.88
FreqNet$_{AAAI'24}$	**100.00**	99.00	99.00	94.50	94.70	99.80	97.10	92.30	76.80	98.90	98.80	99.00	71.60	88.10	89.30	86.40	99.40	93.22
C2P-CLIP$_{AAAI'25}$	**100.00**	99.80	98.40	99.90	**100.00**	**100.00**	**100.00**	98.30	92.30	99.90	**100.00**	99.80	90.90	99.60	99.30	98.60	**100.00**	98.64
D$^3_{CVPR'25}$	99.91	91.85	88.27	98.99	98.67	93.60	99.93	74.93	53.53	87.83	87.69	81.95	82.86	86.15	85.98	85.81	86.54	87.32
VIB$_{CVPR'25}$	**100.00**	97.67	95.90	98.83	99.96	98.97	**100.00**	92.51	89.94	99.30	99.17	93.18	89.78	99.08	96.50	97.37	98.07	96.84
ForgeLens$_{ICCV'25}$	**100.00**	96.80	97.66	99.82	99.77	99.93	99.54	97.29	96.55	99.63	99.74	98.58	92.11	99.54	99.29	98.04	99.60	98.46
LTD$_{CVPR'26}$	**100.00**	99.79	99.04	99.97	**100.00**	**100.00**	99.98	**98.44**	**96.87**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**99.65**
SPARE (Ours)	**100.00**	**99.89**	99.22	99.96	**100.00**	**100.00**	99.94	97.52	95.99	99.88	99.88	99.81	96.47	99.84	99.79	99.72	99.98	99.29

Table 7: AP (%) comparison on DRCT-2M. Each row corresponds to a detection method, and each column to a generative model. The last column reports the mean AP across all models. **Red** and blue indicate the best and second-best results, respectively.

Method	SD variants						Turbo Variants		LCM Variants		ControlNet Variants			Mean AP
	LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	
CNNSpot$_{CVPR'20}$	99.72	99.30	99.22	99.31	94.60	98.47	95.52	97.01	95.94	94.43	95.53	92.87	98.68	96.97
UnivFD$_{CVPR'23}$	99.59	99.25	99.31	99.06	98.66	98.63	97.64	97.47	98.60	98.47	98.66	96.70	97.31	98.41
NPR$_{CVPR'24}$	99.10	**100.00**	**100.00**	96.50	73.10	94.80	**100.00**	**100.00**	**100.00**	92.00	88.20	82.40	90.30	93.57
FatFormer$_{CVPR'24}$	74.24	33.88	33.70	32.95	33.07	38.91	32.99	32.38	35.55	41.07	39.65	43.07	72.36	41.83
RINE$_{ECCV'24}$	99.30	99.40	99.40	99.40	97.90	95.90	96.70	96.70	98.60	97.10	98.80	94.90	97.70	97.83
FreqNet$_{AAAI'24}$	**100.00**	**100.00**	**100.00**	62.00	31.50	38.70	**100.00**	**100.00**	**100.00**	31.50	95.10	93.80	92.40	80.38
C2P-CLIP$_{AAAI'25}$	95.55	98.52	98.51	98.41	96.98	98.30	96.57	96.61	96.82	95.54	97.75	96.53	96.92	97.15
D$^3_{CVPR'25}$	99.00	99.00	99.00	99.00	99.00	99.00	99.00	99.00	99.00	99.00	99.00	99.00	99.00	99.00
VIB$_{CVPR'25}$	98.27	83.26	83.34	89.42	88.92	75.22	84.39	81.00	78.25	93.30	93.00	89.65	93.10	87.01
LTD$_{CVPR'26}$	99.75	**100.00**	**100.00**	99.89	99.86	99.85	**100.00**	**100.00**	**100.00**	99.93	99.97	99.89	99.95	**99.93**
SPARE (Ours)	99.52	**100.00**	**100.00**	**99.93**	**99.94**	**99.91**	**100.00**	**100.00**	**100.00**	**99.96**	**99.98**	**99.92**	**99.96**	**99.93**

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.

Chenming Zhou, Jiaan Wang, Yu Li, Lei Li, Juan Cao, and Sheng Tang. Beyond semantic features: Pixel-level mapping for generalized AI-generated image detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 36101–36109, 2026. doi: 10.1609/aaai.v40i42.40927.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in neural information processing systems*, 36:77771–77782, 2023.

A Appendix

A.1 AP (%) Comparison Against State-of-the-Art

We further report average precision (AP) on ForenSynths, GenImage, and DRCT-2M as a complementary evaluation. Table 6 presents the results on ForenSynths. Following the same two-subclass training protocol as the strongest baseline, LTD, our approach achieves comparable mean AP performance, further demonstrating its effectiveness in distinguishing real from generated images. Table 7 reports the results on DRCT-2M. Our method achieves the same mean AP as LTD, indicating strong robustness to generation dynamics. Table 8 further evaluates cross-dataset generalization on GenImage, where models trained on ForenSynths are directly evaluated on GenImage without additional fine-tuning. Our method achieves performance comparable to LTD, demonstrating strong generalization to unseen data distributions and generation models.

Table 8: AP (%) comparison on GemImage. Each row corresponds to a detection method, and each column to a generative model. The last column reports the mean AP across all models. **Red** and blue indicate the best and second-best results, respectively.

Method	ADM	Midjourney	GLID	VQDM	SDv1.4	SDv1.5	Wukong	BigGAN	Mean AP
CNNSpot$_{CVPR'20}$	61.53	62.67	64.91	56.50	55.04	55.74	55.42	63.19	59.38
UnivFD$_{CVPR'23}$	86.37	74.76	85.34	97.11	86.79	85.07	90.99	98.21	88.08
NPR$_{CVPR'24}$	83.60	**93.70**	93.60	70.60	85.70	84.80	77.80	97.00	85.85
FatFormer$_{CVPR'24}$	91.76	62.77	95.99	97.00	81.12	81.11	85.88	99.55	86.90
RINE$_{ECCV'24}$	93.80	85.10	94.90	98.40	97.50	97.20	97.70	99.20	95.48
FreqNet$_{AAAI'24}$	88.80	69.60	69.30	82.40	58.20	58.70	55.00	88.40	71.30
C2P-CLIP$_{AAAI'25}$	90.12	81.26	94.67	98.30	98.95	98.88	98.60	99.85	95.08
D^3$_{CVPR'25}$	89.04	50.42	81.55	95.78	61.92	61.92	64.31	98.01	75.37
VIB$_{CVPR'25}$	86.66	77.81	82.79	94.82	87.66	86.13	92.75	97.87	88.31
ForgeLens$_{ICCV'25}$	93.00	75.68	96.64	96.56	98.01	98.13	96.54	**99.88**	94.31
LTD$_{CVPR'26}$	**97.65**	82.14	**99.65**	98.99	99.31	**99.19**	99.25	99.80	**97.00**
SPARE (Ours)	97.36	81.23	99.41	**99.50**	**99.36**	98.74	**99.37**	99.63	96.83

Table 9: Patch Shuffle randomness analysis under five independent random permutations. We report mean $\pm$ standard deviation of ACC and AP across these runs.

Dataset	ACC (mean $\pm$ std)	AP (mean $\pm$ std)
ForenSynths	95.76 ± 1.13	99.33 ± 0.02
GenImage	91.09 ± 0.66	96.81 ± 0.06
Average	93.42 ± 0.78	98.07 ± 0.03

Table 10: Comparison of parameter efficiency and inference time. Inference time is measured per image with an input resolution of 224×224.

Method	Parameters (M)		Inference Time (ms)
	All	Trainable	
LTD	440.23	12.61	10
ForgeLens	441.81	14.19	12
SPARE (Ours)	440.23	12.62	15

A.2 Effect of Patch Shuffle Randomness.

SPARE employs Patch Shuffle to disrupt the global spatial structure of an image. Since the patch permutation is stochastic, we further investigate whether its performance is sensitive to different random permutations. Following the default setting, the model is trained with one randomly sampled permutation for each image. During inference, we conduct five independent evaluations using different random permutations, while keeping all other settings unchanged. We report the mean and standard deviation of ACC and AP over these five runs.

As shown in Table 9, SPARE maintains consistently strong performance across different random permutations. In particular, the AP standard deviations are only 0.02% and 0.06% on ForenSynths and GenImage, respectively, resulting in an average AP of $98.07 \pm 0.03\%$. Although ACC exhibits slightly larger variations, its standard deviations remain limited to 1.13% on ForenSynths and 0.66% on GenImage, with an average of $93.42 \pm 0.78\%$. These results demonstrate that SPARE is largely insensitive to the randomness of Patch Shuffle, confirming that the learned structural response captures stable and generalizable cues rather than patterns associated with a particular random permutation.

A.3 Parameter efficiency and Inference Speed

Table 10 compares the parameter efficiency and inference speed of SPARE with representative methods. Built upon CLIP ViT-L/14, SPARE contains 440.23M parameters, of which only 12.62M are trainable, accounting for approximately 2.87% of the total. Its trainable parameter count is comparable to that of LTD and lower than that of ForgeLens, demonstrating that SPARE introduces only lightweight response-learning modules on top of the frozen visual encoder. We further measure the inference time per image using input image of size 224×224. As shown, SPARE requires 15 ms per image, compared with 10 ms for LTD and 12 ms for ForgeLens. This modest increase mainly arises from processing both the original image and its patch-shuffled counterpart through the frozen visual encoder. Nevertheless, SPARE achieves substantial performance improvements while maintaining practical inference efficiency.