

Explainability-guided Deepfake Detection for High-Fidelity Facial Edits

Bibek Das¹, Soumi Chattopadhyay², Chandranath Adak¹, Astitva Pandey¹,
Ashutosh Parihar², Zahid Akhtar³, Soumya Dutta⁴, and Abdenour Hadid⁵

¹ Dept. of CSE, Indian Institute of Technology Patna, India–801106.

² Dept. of CSE, Indian Institute of Technology Indore, India–453552

³ State University of New York Polytechnic Institute, USA–13502

⁴ Dept. of CSE, Indian Institute of Technology Kanpur, India–208016

⁵ SCAI, Sorbonne University Abu Dhabi, UAE

{bibek_2422cs27,chandranath}@iitp.ac.in, soumi@iiti.ac.in

Abstract. Recent advances in vision–language models (VLMs) have enabled highly realistic, identity-preserving facial edits that lack the visible artifacts exploited by conventional deepfake detectors. This raises a critical question for multimedia forensics: do models capable of generating such content also possess reliable forensic reasoning? We first show that state-of-the-art VLMs, despite strong performance on standard benchmarks, fail to consistently detect high-fidelity semantic edits, exhibiting viewpoint sensitivity, prompt-dependent behavior, and reliance on superficial cues rather than manipulation-aware visual evidence. To rigorously study this setting, we introduce IndicSideFace-V2, a multi-view deepfake dataset comprising minimal facial edits with pixel-level manipulation masks, enabling joint evaluation of detection accuracy and explanation fidelity. Our analysis further reveals that standard deep learning detectors, although achieving high accuracy after fine-tuning, frequently succeed for the wrong reasons, attending to background or incidental regions instead of the manipulated facial attributes. Motivated by these findings, we propose an explainability-guided training framework that enforces right prediction for the right reason by explicitly aligning class activation maps with ground-truth manipulation masks during learning. Extensive experiments demonstrate that the proposed approach significantly improves detection robustness and explanation faithfulness under challenging viewpoints and common image perturbations. Together, our results highlight the limitations of language-driven forensic reasoning and establish explanation-aligned visual learning as a necessary step toward trustworthy deepfake detection in the era of high-fidelity generative models.⁶

Keywords: Deepfake Detection · Explainable Artificial Intelligence · Right-for-the-Right-Reasons Learning · High-Fidelity Semantic Facial Editing

⁶ Codes will be made available upon acceptance of the paper. Institutional ethical clearance for conducting this research was obtained from IIT Patna.

1 Introduction

The rapid progress of generative artificial intelligence has fundamentally altered how visual content is created, edited, and disseminated. Modern image generation and editing systems [1] can produce highly photorealistic facial images while preserving identity, pose, illumination, and background, enabling subtle semantic modifications such as changes in expression, hairstyle, or accessories. While these capabilities benefit creative and commercial applications, they also significantly intensify the threat posed by deepfakes to media authenticity, digital trust, and forensic reliability. Unlike earlier face manipulation techniques [2] that introduced visible blending artifacts or geometric inconsistencies, contemporary high-fidelity edits are perceptually natural and difficult to distinguish even for human observers. This shift demands deepfake detection systems that go beyond exploiting superficial artifacts and instead rely on semantically meaningful and spatially grounded evidence.

Large language models (LLMs) [3, 4] and vision-language models (VLMs) [5] have recently emerged as powerful multimodal systems capable of jointly processing visual and textual information. Their impressive generative abilities, coupled with natural-language reasoning interfaces, have prompted growing interest in using them as general-purpose forensic detectors. Intuitively, models that can generate or describe realistic images might be expected to recognize manipulated content as well. However, emerging evidence suggests that this assumption is fragile. Despite occasional high accuracy on legacy benchmarks, VLM-based detectors exhibit unstable behavior under realistic conditions, with predictions that are highly sensitive to prompt formulation, viewpoint variation, and superficial cues such as watermarks or texture heuristics. More critically, their accompanying explanations are often hallucinatory or unverifiable, raising concerns about the reliability of language-driven reasoning for visual forensics.

Recent studies, such as Shan et al. [3], have benchmarked multimodal large language models on standard deepfake datasets, including FaceForensics++. However, these benchmarks primarily focus on identity replacement scenarios (e.g., face swapping) that introduce visible stitching artifacts and low-level inconsistencies. They largely overlook facial attribute editing, where modern vision-language models generate subtle, holistic, and identity-preserving semantic edits without explicit blending boundaries. As a result, existing evaluations fail to reflect the challenges posed by contemporary generative models and may substantially overestimate forensic robustness. Our work fills this gap by evaluating deepfake detection under high-fidelity semantic editing settings that standard benchmarks do not represent.

Traditional deepfake detection approaches [6, 7] predominantly rely on supervised deep learning models trained to classify images or videos as real or manipulated. While such detectors achieve strong performance on curated benchmarks, they often exploit dataset-specific artifacts or background correlations that do not generalize to newer manipulation paradigms. Recent interpretability studies [8, 9] have shown that many detectors are right for the wrong reasons, producing correct predictions while attending to irrelevant regions such as backgrounds

or high-contrast edges rather than the manipulated facial attributes. This limitation is particularly problematic in forensic and legal settings, where decisions must be justified by manipulation-relevant evidence. As generative models increasingly produce identity-preserving and artifact-free edits, the lack of faithful visual reasoning, not merely accuracy, has become a central bottleneck for trustworthy deepfake detection.

In this work, we argue that the failure of vision-language models under high-fidelity semantic edits exposes a deeper limitation in current forensic reasoning: the absence of spatially grounded, manipulation-aware evidence. Rather than proposing a new detector architecture, we take a diagnostic-first approach. We first construct a realistic evaluation setting that reflects modern generative capabilities and systematically analyze the behavior of VLM-based detectors under minimal-edit conditions. Motivated by the observed shortcomings, we demonstrate that standard convolutional detectors can recover robustness and reliability when explicitly constrained to base their decisions on the manipulated regions. To this end, we introduce an explainability-guided training framework that enforces right prediction for the right reason by aligning model attention with ground-truth manipulation masks during learning. The main contributions of this work are summarized below:

- (i) *Forensic Analysis of Vision-Language Models*: We provide a systematic analysis showing that state-of-the-art VLMs fail to reliably detect minimal semantic edits, exhibiting prompt sensitivity, viewpoint bias, and reliance on superficial heuristics rather than manipulation-aware reasoning.
- (ii) *Explainability-Guided Deepfake Detection*: We propose an explanation-aligned training objective that enforces right-for-the-right-reasons learning, significantly improving robustness and explanation faithfulness of deepfake detectors under realistic conditions.
- (iii) *Dataset and Evaluation Setting*: We introduce IndicSideFace-V2, a challenging multi-view deepfake dataset featuring identity-preserving, high-fidelity facial edits with pixel-level manipulation masks, enabling joint evaluation of detection accuracy and explanation fidelity.

2 Proposed Dataset: IndicSideFace-V2

This section introduces IndicSideFace-V2, a dataset designed to evaluate the realism of VLMs-driven image generation under high-fidelity, minimal-manipulation settings. Recent advances in generative models enable fine-grained semantic edits while preserving identity, pose, illumination, and background. IndicSideFace-V2 reflects this capability through real-fake image pairs that are visually natural and differ only by subtle, localized modifications, resulting in a challenging and realistic detection setting.

2.1 Genuine Data Collection & VLMs-based Fake Image Generation

To capture demographic diversity, we selected 140 individuals of Indian origin from different regions, covering a wide range of skin tones and facial character-

istics. This selection aims to represent diverse ethnic and cultural backgrounds. The dataset includes 26 female and 114 male subjects, with 40 individuals wearing glasses.

IndicSideFace-V2 comprises 1680 images, including 840 genuine images, 840 synthetic fake images, and corresponding 840 photometric frequency-guided ground-truth masks. Motivated by [10], we collected images under six controlled viewpoints to disentangle manipulation cues from pose and environmental factors for each subject: Indoor-Front (IF), Indoor-Left (IL), Indoor-Right (IR), Outdoor-Front (OF), Outdoor-Left (OL), and Outdoor-Right (OR).

Explicit control over viewpoints enforces balanced coverage across acquisition conditions and prevents detectors from exploiting spurious correlations. A uniform split of 8:1:1 is applied per viewpoint for both genuine and fake data classes. The dataset supports multiple evaluation tasks, including binary classification, pixel-level manipulation localization, and explanation alignment, using a subject and viewpoint-stratified protocol.

Genuine images are collected under diverse environmental settings using consumer-grade smartphone cameras, capturing natural variations in background, illumination, and appearance. Synthetic fake images are then generated from the corresponding genuine images by 140 unique users from diverse regions of India, each applying image editing through the ‘Gemini Nano Banana’ [11] model using user-defined natural language prompts. These prompts reflect diverse real-world user intent, producing localized and semantically meaningful edits spanning facial expressions, appearance attributes, and accessories, while preserving identity, pose, lighting, and background consistency. This human-in-the-loop image generation process reflects real-world manipulation practices and produces highly realistic fake images with minimal perceptual changes. Each fake image is lightly refined using basic post-processing operations to enhance visual naturalness. Pixel-level ground-truth masks are then generated by aligning the refined

	Indoor-Left	Indoor-Front	Indoor-Right	Outdoor-Left	Outdoor-Front	Outdoor-Right
Real						
Prompt	"change expression to smile."	"make him wear specs."	"Give Long hair."	"Give a big beard."	"change tshirt to formal shirt."	"make him wear big gold chain and white cap"
Fake						

Fig. 1. Representative examples from IndicSideFace-V2 showing real images, user-defined editing prompts, and generated fake images across indoor/ outdoor viewpoints.

fake image with its real counterpart, followed by conservative pixel-wise difference thresholding and manual verification to eliminate false positives caused by illumination or alignment variations.

Fig. 1 presents representative examples from IndicSideFace-V2, showing real images, user-define prompts, and the corresponding generated fake images across different indoor and outdoor viewpoints.

2.2 Realism Validation and Evaluation Protocols

To assess perceptual realism, we conducted a controlled human study in which participants manually marked 50 random samples of IndicSideFace-V2 as real or fake. The average score of correct identification was 25.19 out of 50, with a median score of 25, corresponding to near-random performance. This result indicates that the VLM-generated fake images are difficult to distinguish from real images at a perceptual level.

We further evaluate realism using objective image-quality metrics, where the BRISQUE score distributions for real and fake images substantially overlap, with a mean difference of 2.40. Additionally, the Fréchet Inception Distance (FID) between real and fake image sets is 62.16, indicating preservation of natural image statistics [12, 13]. Fig. 2 visualizes the BRISQUE distributions and reports the mean BRISQUE and FID values.

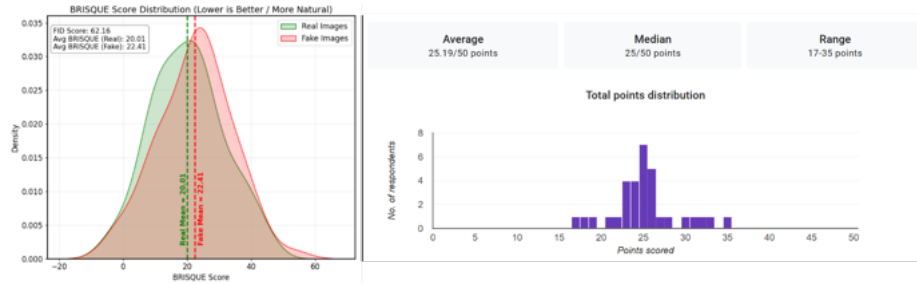


Fig. 2. BRISQUE score distributions for real and fake images. Dashed lines indicate class means; the inset reports mean BRISQUE and FID values.

3 VLMs as Forensic Detectors

To assess whether recent advances in generative and multimodal models translate into reliable forensic reasoning, we next use IndicSideFace-V2 as a controlled testbed for deepfake detection. Unlike existing benchmarks, this dataset isolates identity-preserving, high-fidelity semantic edits under systematic viewpoint variation, thereby eliminating spurious shortcuts related to pose, background, or visible blending artifacts. This setting enables a rigorous evaluation of whether

Table 1. Detection Performance of VLMs across Different Datasets and Attack Types

Dataset /	ChatGPT 5.1 [3]				Gemini 2.5 Flash [11]				Gemini 3 Pro [14]				Grok 4.1 [15]			
Attack Type	$\mathcal{A} \uparrow$	$F_1 \uparrow$	$TPR \uparrow$	$FPR \downarrow$	$\mathcal{A} \uparrow$	$F_1 \uparrow$	$TPR \uparrow$	$FPR \downarrow$	$\mathcal{A} \uparrow$	$F_1 \uparrow$	$TPR \uparrow$	$FPR \downarrow$	$\mathcal{A} \uparrow$	$F_1 \uparrow$	$TPR \uparrow$	$FPR \downarrow$
IndicSideFace-V1 (Front)	70.0	57.1	40.0	0.0	50.0	0.0	0.0	0.0	50.0	0.0	0.0	0.0	40.0	0.0	0.0	20.0
IndicSideFace-V1 (Left)	70.0	57.1	40.0	0.0	55.0	18.2	10.0	0.0	50.0	0.0	0.0	0.0	35.0	0.0	0.0	30.0
IndicSideFace-V1 (Right)	70.0	57.1	40.0	0.0	60.0	33.3	20.0	0.0	50.0	0.0	0.0	0.0	45.0	0.0	0.0	10.0
IndicSideFace-V1 (mean)	70.0	57.1	40.0	0.0	55.0	17.2	10.0	0.0	50.0	0.0	0.0	0.0	40.0	0.0	0.0	20.0
IndicSideFace-V2 (Front)	66.7	48.0	31.6	0.0	50.0	0.0	0.0	0.0	55.0	18.2	10.0	0.0	90.0	88.9	80.0	0.0
IndicSideFace-V2 (Left)	65.0	46.2	30.0	0.0	50.0	0.0	0.0	0.0	77.5	71.0	55.0	0.0	90.0	88.9	80.0	0.0
IndicSideFace-V2 (Right)	95.0	94.7	90.0	0.0	52.5	9.5	5.0	0.0	67.5	51.8	35.0	0.0	82.5	78.8	65.0	0.0
IndicSideFace-V2 (mean)	75.8	68.1	51.7	0.0	50.8	3.3	1.7	0.0	66.7	50.0	33.3	0.0	87.5	85.7	75.0	0.0
Moiré Pattern	50.8	3.3	1.7	0.0	50.0	0.0	0.0	0.0	66.7	60.0	50.0	16.7	41.7	0.0	0.0	16.7
Brightness	52.5	9.5	5.0	0.0	51.7	6.5	3.3	0.0	64.2	44.2	28.3	0.0	46.7	0.0	0.0	6.7
Contrast	52.5	9.5	5.0	0.0	51.7	6.5	3.3	0.0	65.0	46.2	30.0	0.0	50.0	0.0	0.0	0.0
Saturation	51.7	6.5	3.3	0.0	50.0	0.0	0.0	0.0	64.2	44.2	28.3	0.0	54.2	26.7	16.7	8.3
Erase	43.3	5.6	3.3	16.7	62.5	43.0	28.3	3.3	62.5	40.0	25.0	0.0	66.7	50.0	33.3	0.0
Gaussian Blur	51.7	6.5	3.3	0.0	50.0	0.0	0.0	0.0	65.0	46.2	30.0	0.0	50.0	0.0	0.0	0.0
Gaussian Noise	51.7	6.5	3.3	0.0	50.8	3.3	1.7	0.0	55.8	39.1	28.3	16.7	50.0	0.0	0.0	0.0
JPEG	50.8	3.3	1.7	0.0	50.8	3.3	1.7	0.0	61.7	37.8	23.3	0.0	50.0	0.0	0.0	0.0
Instagram	57.5	26.1	15.0	0.0	50.0	0.0	0.0	0.0	64.2	44.2	28.3	0.0	50.0	0.0	0.0	0.0
White Balance	51.7	6.5	3.3	0.0	50.0	0.0	0.0	0.0	64.2	44.2	28.3	0.0	46.7	0.0	0.0	6.7

$\mathcal{A}$: Accuracy, F_1 : F1 score, TPR : True positive rate, FPR : False positive rate

detectors genuinely respond to manipulation-aware visual evidence or rely on superficial cues. In the following section, we leverage this benchmark to systematically analyze the behavior of state-of-the-art vision-language models as forensic detectors, examining their accuracy, robustness, and consistency under minimal-edit conditions.

To examine whether generative VLMs possess reliable forensic reasoning capabilities, we evaluate several state-of-the-art large multimodal models, including ChatGPT 5.1 [3], Gemini 2.5 Flash [11], Gemini 3 Pro [14], and Grok 4.1 [15], under diverse deepfake detection scenarios. Table 1 details their performance on face-swap datasets specifically (IndicSideFace [10] V1 and V2) and analyzes their robustness under a wide array of photometric and compression-based perturbations.

Performance on High-Fidelity Edits: A noticeable observation from Table 1 is the sharp discrepancy in model reliability when faced with realistic semantic edits. On IndicSideFace-V2 (mean), ChatGPT 5.1 achieves $\mathcal{A} = 75.8$ but with a modest $TPR = 51.7$, while Gemini 2.5 Flash collapses to near-chance performance ($\mathcal{A} = 50.8$, $TPR = 1.7$). These results indicate that high reported accuracy in certain scenarios often masks degenerate prediction strategies, such as a majority real-class bias, rather than reflecting genuine sensitivity to sophisticated manipulations.

Viewpoint Sensitivity and Degenerate Recall: Table 1 further revealed pronounced viewpoint sensitivity. Across IndicSideFace-V1 and V2, detection performance degrades significantly for left and right profile views. In multiple cases, models exhibit near-zero recall for side views despite reasonable frontal performance. For example, Gemini 3 Pro shows TPR zero for both Left, Right and front views in IndicSideFace-V1. Such behavior suggests that VLMs fail to model localized, pose-dependent manipulation cues, instead relying on global appearance statistics that break under controlled viewpoint variation.

Brittleness under Mild Perturbations: Robust forensic detectors must tolerate benign post-processing operations common in real-world media pipelines. However, Table 1 demonstrates that VLM-based detectors are extremely brittle under even mild perturbations. Across brightness, contrast, saturation, Gaussian



Fig. 3. Qualitative failure analysis of VLMs on IndicSideFace-V2. Under minimal, identity-preserving edits, VLMs produce contradictory real/fake predictions and inconsistent explanations across prompts and roles. Decisions are driven primarily by prompt-sensitive heuristics and watermark detection (e.g., SynthID [16]) rather than localized manipulation cues, revealing hallucinated reasoning.

blur, Gaussian noise, JPEG compression, and social-media style filtering, most models' $\mathcal{F}_1$ scores collapse toward zero, with true positive rates frequently below 50%. Notably, this degradation occurs even when false positive rates remain near zero, further confirming a conservative real-class bias rather than stable manipulation awareness.

Qualitative Evidence of Heuristic-Driven Reasoning: The qualitative examples in Fig. 3 provide further insight into the underlying failure modes. Under near-identical visual conditions, VLMs often produce contradictory real/fake verdicts across different prompts, accompanied by confident but unverifiable explanations. In several cases, models rely on external cues such as watermark detection (e.g., SynthID [16]) rather than intrinsic visual inconsistencies. Simple post-processing operations that suppress watermark visibility are sufficient to reverse model decisions, demonstrating that such verification mechanisms are watermark-dependent rather than manipulation-aware. Moreover, explanations frequently cite vague concepts such as *unnatural texture* or *inconsistent lighting* without consistent localization, exposing hallucinated forensic reasoning.

Collectively, the results in Table 1 and Fig. 3 expose a fundamental forensic-generative gap in contemporary vision-language models. Although such models excel at producing highly realistic facial edits, their detection behavior is inconsistent, viewpoint-dependent, and non-robust to minor perturbations. Crucially, high accuracy on standard datasets does not correspond to faithful visual

reasoning; instead, decisions appear driven by prompt-sensitive or watermark-related cues rather than spatially grounded manipulation evidence.

These limitations motivate a shift away from language-driven forensic reasoning toward explicitly grounded visual learning paradigms. The failure modes observed in Table 1 suggest that reliable deepfake detection requires models to localize and reason over manipulated regions rather than infer authenticity from global or semantic heuristics. In the following sections, we therefore investigate traditional convolutional detectors and show that, while they achieve higher quantitative accuracy, they too often succeed for the wrong reasons. This observation directly motivates our proposed explainability-guided training framework, which enforces right prediction for the right reason by aligning model attention with ground-truth manipulation masks, thereby enabling robust and forensically meaningful deepfake detection under high-fidelity semantic edits.

4 Traditional DL Models as Forensic Detectors

While Section 3 demonstrates that VLM-based detectors fail under realistic, minimally manipulated conditions due to unreliable semantic reasoning, traditional DL models/ CNN-based detectors achieve stronger quantitative performance after task-specific fine-tuning. However, despite improved accuracy and $\mathcal{F}_1$ scores (refer to Table 2), their forensic reliability remains limited. As illustrated in Fig. 4, importance maps show that, despite correct predictions, CNN decisions are often guided by spatially misaligned activations concentrated on background textures or incidental facial regions rather than the true manipulated areas. This behavior reflects shortcut learning and exemplifies the Right for the Wrong Reasons phenomenon, where numerical correctness is achieved without grounding decisions in manipulation-relevant visual evidence, undermining the suitability of standard CNN detectors for explainable forensic analysis.

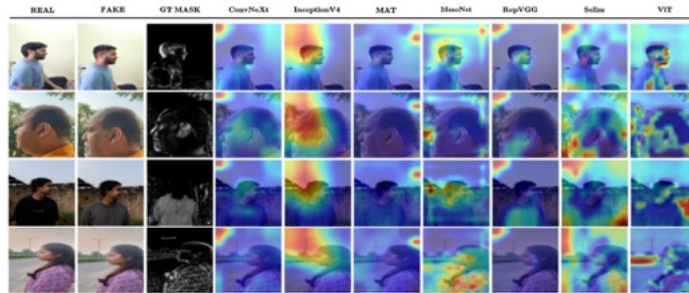


Fig. 4. Explainability limitation of CNN-based detectors: despite correct fake classification, importance maps are partially misaligned with the manipulated regions.

Table 2. Quantitative comparison of classification ($\mathcal{A}$, $\mathcal{F}_1$, $\mathcal{AUC}$) and localization ($\mathcal{IoU}_{0.5}$, $m\mathcal{IoU}$, $\mathcal{ELM}_{\text{soft}}$, $\mathcal{CAM}$) metrics on IndicSideFace-V2.

Type	Model	$\mathcal{A} \uparrow$	$\mathcal{F}_1 \uparrow$	$\mathcal{AUC} \uparrow$	$\mathcal{IoU}_{0.5} \uparrow$	$m\mathcal{IoU} \uparrow$	$\mathcal{ELM}_{\text{soft}} \uparrow$	$\mathcal{CAM}\text{-Score} \uparrow$
Deep Learning	InceptionV4	95.83	96.00	99.94	56.87	38.76	11.02	10.41
	ConvNeXt-B	95.00	94.74	99.44	64.71	44.84	9.63	8.96
	XceptionNet	94.17	94.40	99.42	61.68	41.54	11.49	10.87
	InceptionResNetV2	91.67	92.19	99.11	51.43	34.45	11.54	10.71
	DeiT-B-16	80.00	83.33	98.89	69.30	48.59	10.40	9.86
	RepVGG-B3	90.83	91.47	98.78	33.69	16.74	11.30	9.30
	RegNetY-32G	95.83	95.80	98.64	32.55	19.36	11.18	8.96
	DenseNet201	85.00	86.76	97.00	32.33	18.10	10.90	9.65
	EfficientNetB4	89.17	88.70	96.92	24.64	7.44	9.82	8.68
	MobileNetV3-L	85.00	85.48	94.11	42.37	23.46	11.94	10.73
	MixNetXL	88.33	88.71	93.90	66.53	44.01	10.08	9.48
	ResNet50V2	80.00	75.51	92.81	19.04	10.05	13.05	10.53
	ResNeXt50_32x4d	85.83	85.71	91.72	23.39	12.73	13.22	10.81
	ViT-B	50.83	42.72	61.83	65.88	50.62	9.56	8.81
Forensic	Selim [17]	77.50	70.97	97.75	69.49	57.92	9.64	9.19
	MAT [18]	85.83	85.22	94.67	25.66	7.62	9.75	8.69
	MesoNet [6]	50.00	53.85	51.72	69.36	47.18	11.81	10.99
	CapsuleNet [19]	51.67	43.14	49.78	69.50	57.14	9.74	9.17
	Proposed	97.50	97.44	99.94	70.92	65.21	11.69	11.30

5 Proposed Explainability-guided Model

The analysis in Sections 3 and 4 reveals a critical limitation of existing deepfake detectors: high classification accuracy does not necessarily imply correct forensic reasoning. Vision-language models fail due to weak spatial grounding, while conventional CNN-based detectors, despite stronger inductive biases, frequently rely on shortcut cues such as background textures, illumination gradients, or facial contours. As demonstrated by $\mathcal{CAM}$ visualizations (Fig. 4), these models often succeed for the *wrong reasons*, attending to regions unrelated to the manipulated facial attributes. For forensic applications, such behavior is unacceptable. A reliable detector must not only predict whether an image is fake, but must do so by attending to the manipulated regions themselves. This motivates a training paradigm that explicitly constrains *where* a model is allowed to derive evidence from, rather than relying solely on label supervision.

Model Formulation: Let $x \in \mathbb{R}^{H \times W \times C}$ denote an input facial image, where H , W , and C represent the image height, width, and number of channels, respectively. Let $y \in \{0, 1\}$ denote the corresponding binary label indicating real or fake, and let $M \in \{0, 1\}^{H \times W}$ be the ground-truth manipulation mask identifying the edited pixels. We denote a CNN-based deepfake detector by $f_\theta(\cdot)$, parameterized by learnable weights θ . Given an input image x , the model produces a prediction $\hat{y} = f_\theta(x)$ and a normalized mean feature activation map $\hat{M}_\theta(x) \in [0, 1]^{H \times W}$, obtained by channel-wise averaging of the final feature representations, which captures the spatial distribution of discriminative features contributing to the model’s decision.

Objective: To enforce *Right Prediction for the Right Reason*, we optimize a joint objective that combines classification accuracy with explanation alignment:

$$\min_{\theta} \mathbb{E}_{(x,y,M)} \left[\mathcal{L}_{\text{cls}}(f_{\theta}(x), y) + \lambda \mathcal{L}_{\text{mask}}(\hat{M}_{\theta}(x), M) \right], \quad (1)$$

where, $\mathcal{L}_{\text{cls}}$ denotes the binary cross-entropy loss, $\mathcal{L}_{\text{mask}}$ enforces spatial consistency between the explanation map and the manipulation mask, and $\lambda > 0$ controls the strength of explanation supervision.

Mask Alignment Loss: The explanation alignment loss is defined as a weighted combination of region overlap and spatial smoothness constraints:

$$\mathcal{L}_{\text{mask}}(\hat{M}, M) = w_{\text{IoU}} \mathcal{L}_{\text{IoU}}(\hat{M}, M) + w_{\text{TV}} \mathcal{L}_{\text{TV}}(\hat{M}), \quad (2)$$

where, w_{IoU} and w_{TV} are scalar weighting coefficients.

Soft IoU Loss: The soft Intersection-over-Union loss encourages maximal spatial overlap between the activation map and the manipulation mask:

$$\mathcal{L}_{\text{IoU}}(\hat{M}, M) = 1 - \frac{\sum_i \hat{M}_i M_i + \varepsilon}{\sum_i \hat{M}_i + \sum_i M_i - \sum_i \hat{M}_i M_i + \varepsilon} \quad (3)$$

where, i indexes spatial pixel locations and ε is a small constant introduced for numerical stability.

Total Variation Regularization: To suppress noisy or fragmented explanations while preserving meaningful boundaries, we add a total variation regularizer:

$$\mathcal{L}_{\text{TV}}(\hat{M}) = \sum_{i,j} \left(|\hat{M}_{i+1,j} - \hat{M}_{i,j}| + |\hat{M}_{i,j+1} - \hat{M}_{i,j}| \right), \quad (4)$$

which promotes spatial smoothness in the explanation map.

The classification loss $\mathcal{L}_{\text{cls}}$ ensures label correctness, while the explanation loss $\mathcal{L}_{\text{mask}}$ explicitly constrains the spatial evidence used by the network. Importantly, the CNN architecture remains unchanged from the baseline; the same backbone is retrained using the proposed objective. Consequently, the observed improvements arise solely from explanation-level supervision rather than architectural modifications. By explicitly aligning model attention with ground-truth manipulation regions, the proposed framework enforces correct predictions through correct reasoning, yielding detectors that are accurate, robust, and forensically interpretable.

Network Architecture and Mask Construction: In our implementation, the detector $f_{\theta}(\cdot)$ is instantiated using a ConvNeXt-B backbone, chosen for its strong representational capacity and stable spatial inductive biases. Given an input image $x \in \mathbb{R}^{H \times W \times C}$, ConvNeXt-B extracts a hierarchy of convolutional feature maps. Let $F_{\theta}(x) \in \mathbb{R}^{h \times w \times d}$ denote the output of the final convolutional stage, where (h, w) is the spatial resolution and d is the channel dimensionality. Global average pooling followed by a fully connected classifier $g(\cdot)$ produces the final prediction, $\hat{y} = g(\text{GAP}(F_{\theta}(x)))$, where $\text{GAP}(\cdot)$ denotes global average pooling.

The class activation map used for explanation alignment is derived directly from the last convolutional layer. Specifically, the predicted mask $\hat{M}_\theta(x) \in [0, 1]^{H \times W}$ is obtained by linearly combining the feature maps in $F_\theta(x)$ using the classifier weights corresponding to the predicted class, followed by bilinear upsampling and min-max normalization to the input resolution.

The ground-truth manipulation mask M is constructed by fusing complementary pixel-level and frequency-domain cues between a real image x^r and its manipulated counterpart x^f . Formally, we define:

$$M = \alpha_1 |x^r - x^f| + \alpha_2 (1 - \text{SSIM}(x^r, x^f)) + \alpha_3 |\mathcal{F}(x^r) - \mathcal{F}(x^f)|, \quad (5)$$

where $\text{SSIM}(\cdot, \cdot)$ denotes the structural similarity index computed locally over the image, $\mathcal{F}(\cdot)$ represents the Fourier transform capturing absolute frequency differences, and $\alpha_1, \alpha_2, \alpha_3 \geq 0$ are weighting coefficients satisfying $\alpha_1 + \alpha_2 + \alpha_3 = 1$. The resulting mask is thresholded and refined to suppress spurious responses caused by illumination or alignment noise.

This formulation yields a manipulation mask that captures both spatial appearance changes and frequency-domain inconsistencies, providing reliable pixel-level supervision for enforcing explanation alignment during training.

6 Experimental Analysis

Experimental Setup: All experiments are conducted on IndicSideFace-V2 using a subject and viewpoint stratified 9:1 train/validation split, with evaluation conducted on an independent test set. We evaluate three categories of models: (i) general-purpose deep learning classifiers (e.g., ConvNeXt-B, XceptionNet, Inception variants, DenseNet, EfficientNet, RegNetY, RepVGG, MixNet, ResNeXt, and Vision Transformers like ViT and DeiT), (ii) forensic-specific detectors (MAT, Selim, MesoNet, CapsuleNet), and (iii) the proposed explainability-guided model based on ConvNeXt-B. All baselines are fine-tuned under identical protocols for fair comparison.

Ground-truth manipulation masks are generated using Eq. 5 with empirically selected weights $\alpha_1 = 0.4$, $\alpha_2 = 0.3$, and $\alpha_3 = 0.3$, balancing appearance, structural, and frequency-domain cues. Evaluation follows both predictive and explanatory criteria. Predictive metrics include Accuracy ($\mathcal{A}$), F1 score ($\mathcal{F}_1$), and Area Under the Curve ($\mathcal{AUC}$). Explainability is assessed using localization metrics: Intersection over Union at threshold 0.5 ($\mathcal{IoU}_{0.5}$), Mean IoU across thresholds 0.5–0.9 ($m\mathcal{IoU}$), Soft-EIM ($\mathcal{EIM}_{\text{soft}}$), and CAM-Score ($\mathcal{CAM}$).

Predictive Performance (Table 2): Several high-capacity backbones achieved strong classification accuracy ($\geq 95\%$) and high $\mathcal{AUC}$ ($> 99\%$). For instance, InceptionV4 matched the proposed method’s $\mathcal{AUC}$ (99.94%) but trailed in other metrics. The proposed model achieved the best overall balance, securing the highest accuracy of 97.50% and $\mathcal{F}_1$ score of 97.44%, while maintaining a state-of-the-art $\mathcal{AUC}$ of 99.94%. In contrast, specialized forensic detectors (e.g., Selim, MAT) generally lagged behind the general-purpose backbones, yielding significantly lower classification scores ($\mathcal{A} < 86\%$) despite their domain-specific design.

Accuracy–Explainability Gap (Table 2): Despite strong predictive performance, many baselines exhibited poor localization quality. For instance, RegNetY-32G achieved high accuracy (95.83%) but failed to accurately localize manipulations, yielding a low $mIoU$ of 19.36%. Similarly, InceptionV4 matched the proposed model’s predictive AUC yet trailed by over 26% in $mIoU$ (38.76% vs. 65.21%). In contrast, the proposed model attained the highest localization fidelity, securing the top scores in $IoU_{0.5}$ (70.92%), $mIoU$ (65.21%), and $CA\mathcal{M}$ (11.30%). While ResNeXt50 achieved marginally higher $\mathcal{EIM}_{\text{soft}}$ (13.22%), it suffered from significantly lower predictive accuracy (85.83%). The proposed method effectively closed the gap between prediction correctness and forensic validity, ensuring that model decisions were grounded in precise manipulation evidence.

Robustness under Adversarial Attacks: Fig. 5 provided an attack-wise robustness and explanation-stability analysis across models, revealing clear qualitative and quantitative distinctions in forensic behavior. The left radar plot ($\mathcal{A}$) depicted that the Proposed model maintained the highest accuracy across almost all perturbations, exceeding 95% under contrast, erase, blur, saturation, and white-balance attacks, while preserving 79.2% accuracy under Gaussian noise, resulting in a significantly more stable performance envelope than all baselines. Despite strong clean performance, high-capacity backbones degraded substantially relative to the Proposed model, exhibiting 15–18% lower accuracy under Gaussian noise (ConvNeXt-B: 64.2%, InceptionV4: 60.8% vs. Proposed: 79.2%) and 6–10% lower accuracy under JPEG compression (85.0%, 81.7% vs. Proposed: 91.6%), indicating reduced robustness to signal corruption, while forensic detectors such as CapsuleNet and MesoNet exhibited a near-total collapse in classification performance, with accuracy frequently falling below 50%. The middle plot (AUC) showed that the Proposed model sustained near-saturated performance across all attacks (AUC : 0.985–0.999), forming a stable outer envelope, whereas CapsuleNet and MesoNet exhibited severely degraded discrimination with $AUC < 0.55$ under multiple perturbations. While forensic-specific detectors (MesoNet, Cap-

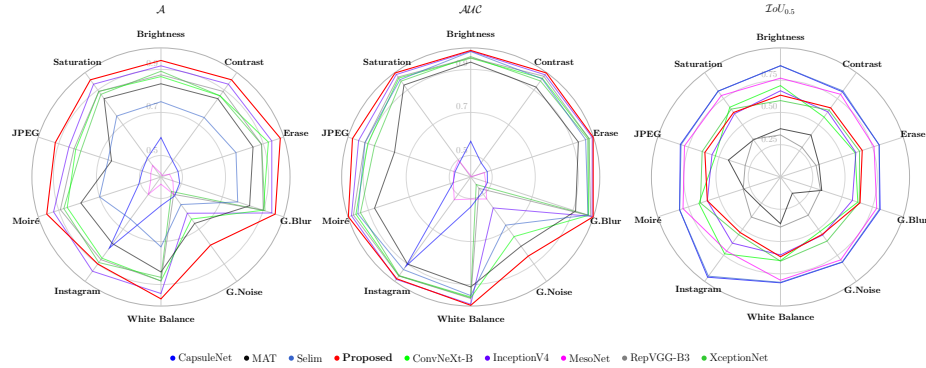


Fig. 5. Performance of different models under adversarial attack conditions.

suleNet, Selim) achieved high $\mathcal{IoU}_{0.5}$ (≈ 0.77 - 0.83), their limited accuracy and $\mathcal{AUC}$ indicated weak integration of spatial cues into global decisions; conversely, under JPEG compression, InceptionV4 and ConvNeXt-B lagged the proposed model by $\mathcal{IoU}_{0.5} \approx 6$ - 9 (e.g., 0.61 vs. 0.55 and 0.52), while the Proposed model consistently preserved high accuracy, strong AUC, and stable spatial grounding, thereby enhancing both interpretability and forensic reliability under adverse conditions.

Generalization across Face-Swap Architectures: (Table 3): Trained on FaceDancer [10] and evaluated zero-shot on Ghost and SimSwap, the Proposed model demonstrates strong cross-architecture generalization. It achieves 95.58% Accuracy on FaceDancer while nearly doubling localization performance over ConvNeXt-B ($\mathcal{IoU}_{0.5} = 35.12\%$). This advantage persists on Ghost, where it attains the highest Accuracy (94.44%) and localization fidelity (37.22%). Although baselines such as XceptionNet remain competitive in classification, they consistently exhibited localization shrinkage, with inferior $\mathcal{EIM}_{\text{soft}}$ and $\mathcal{CAM}$ -Score. Even on the challenging SimSwap dataset, the proposed method maintains a favorable balance between classification reliability (93.86%) and robust feature grounding, indicating that explainability-guided supervision effectively learns manipulation-invariant representations.

Qualitative Analysis of Explanation Fidelity: Fig. 6 provided a qualitative comparison of saliency map visualizations for representative baseline detectors and the proposed explainability-guided model on IndicSideFace-V2, spanning multiple viewpoints and high-fidelity semantic edit types. Despite correct predictions, baseline CNN-based detectors consistently exhibited diffuse and spatially misaligned activations, frequently attending to background regions, facial contours,

Table 3. Quantitative comparison on FaceDancer, Ghost, and SimSwap subsets of the IndicSideFace-V1 [10].

	Model	$\mathcal{A} \uparrow$	$\mathcal{F}_1 \uparrow$	$\mathcal{IoU}_{0.5} \uparrow$	$m\mathcal{IoU} \uparrow$	$\mathcal{EIM}_{\text{soft}} \uparrow$	$\mathcal{CAM}$ -Score $\uparrow$
FaceDancer	Proposed	95.58	95.59	35.12	28.43	9.83	8.96
	ConvNeXt-B	92.92	93.22	16.05	19.89	2.56	2.50
	XceptionNet	94.25	94.20	21.03	19.51	4.59	4.39
	ResNeXt50-32x4d	77.21	81.24	16.97	19.08	4.02	3.90
	ResNet-50	65.27	73.96	7.66	13.52	2.04	2.01
Ghost	Proposed	94.44	94.30	37.22	30.72	11.60	10.56
	ConvNeXt-B	92.95	93.28	21.90	25.44	3.46	3.40
	XceptionNet	91.24	90.83	21.16	21.70	5.42	5.28
	ResNeXt50-32x4d	67.31	75.12	25.32	21.27	6.19	5.92
	ResNet-50	74.36	79.24	9.52	16.27	2.66	2.63
SimSwap	Proposed	93.86	93.81	18.28	14.03	9.24	8.24
	ConvNeXt-B	93.20	93.53	18.47	15.31	3.11	3.02
	XceptionNet	90.57	90.16	27.89	25.97	2.51	2.46
	ResNeXt50-32x4d	76.10	80.07	26.39	22.42	4.69	4.50
	ResNet-50	64.25	73.23	24.41	17.51	5.41	5.10

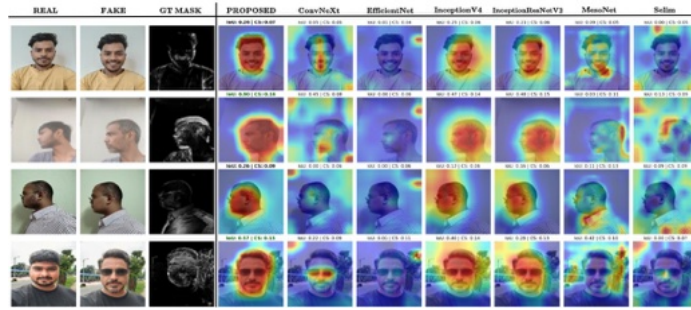


Fig. 6. Qualitative comparison of explanation heatmaps between baseline and proposed models on IndicSideFace-V2

or high-contrast edges rather than the manipulated facial attributes. This behavior is observed even for backbones such as ConvNeXt, EfficientNet, and InceptionResNetV2, which frequently produce activations in irrelevant background areas or large, non-specific regions of the face. In contrast, the Proposed model produces compact, localized, and semantically meaningful activation maps that closely align with the true edited regions across both frontal and side-face views, consistently achieving higher $\mathcal{IoU}_{0.5}$ and $\mathcal{CAM}$ -Score in comparison to all baselines. This qualitative superiority directly reflects the effect of the proposed Right Prediction for the Right Reasons objective, which explicitly constrains model reasoning to be spatially grounded in manipulation-relevant regions during training. Importantly, the visual trends in Fig. 6 are quantitatively corroborated by the explainability metrics in Table 2, where the proposed model achieves the highest $\mathcal{IoU}_{0.5}$ (70.92%), $\mathcal{CAM}$ (11.30%), and $\mathcal{AUC}_{\text{ins}}$ (99.94%), confirming that its salient regions are both spatially aligned and causally responsible for correct predictions. Together, these results demonstrate that while baseline models may achieve high classification accuracy, their reasoning remains unreliable, whereas the proposed model attains correctness through faithful and manipulation-grounded visual explanations, making it suitable for trustworthy forensic deployment.

7 Conclusion

This work demonstrates that high-fidelity, identity-preserving facial edits expose fundamental limitations in both vision-language models and conventional deepfake detectors, motivating the need for manipulation-grounded forensic reasoning. By introducing IndicSideFace-V2 and an explainability-guided Right-for-the-Right-Reasons training framework, we show that aligning model attention with manipulated regions yields simultaneous gains in accuracy, robustness, and explanation fidelity. A current limitation of our approach is its reliance on pixel-level manipulation masks, which may not always be available or scalable in real-world settings. Future work will explore weakly supervised and self-supervised

strategies for explanation alignment, extend the framework to video-based and cross-generator scenarios, and investigate integration with open, reproducible vision-language models to further improve generalization and forensic reliability.

Acknowledgments. Partial support from IHUB-NTIHAC Foundation, IIT Kanpur, under proposal no. 3092 is gratefully acknowledged.

References

1. Ouxiang Li et al. Easier Painting Than Thinking: Can Text-to-Image Models Set the Stage, but Not Direct the Play? *arXiv preprint arXiv:2509.03516*, 2025.
2. Tero Karras et al. Analyzing and Improving the Image Quality of StyleGAN. In *CVPR*, 2020.
3. Shan Jia et al. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *CVPRW*, pages 4324–4333, 2024.
4. Gemini Team et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
5. Siyuan Cheng et al. SHIELD: A Benchmark Study on Zero-Shot Detection of AI-Edited Images with Vision Language Models. In *Generative and Protective AI for Content Creation*, 2025.
6. Darius Afchar et al. Mesonet: a compact facial video forgery detection network. In *WIFS*, pages 1–7. IEEE, 2018.
7. Eshika Jain et al. EfficientNet-Based Deepfake Detection: A Robust Approach for Real and Fake Media Classification. In *GCCIT*, pages 1–6. IEEE, 2024.
8. Andrew Slavin Ross et al. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *IJCAI-17*, pages 2662–2670, 2017.
9. Zhuofan Ying et al. Visfis: Visual feature importance supervision with right-for-the-right-reason objectives. *NeurIPS*, 35:17057–17072, 2022.
10. Anurag Deo et al. IndicSideFace: A Dataset for Advancing Deepfake Detection on Side-Face Perspectives of Indian Subjects. In *IEEE FG*, pages 1–10, 2025.
11. Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
12. Anish Mittal et al. Blind/referenceless image spatial quality evaluator. In *ASISLOMAR*, pages 723–727. IEEE, 2011.
13. Martin Heusel et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NIPS*, 30, 2017.
14. Gemini Team et al. A new era of intelligence with gemini 3, 2025.
15. xAI Team. Grok 4.1 model card, 2025.
16. Sven Gowal et al. SynthID-Image: Image watermarking at internet scale. *arXiv preprint arXiv:2510.09263*, 2025.
17. Selim Seferbekov. DeepFake Detection (DFDC) Solution. https://github.com/selimsef/dfdc_deepfake_challenge, Accessed: 11 June 2025.
18. Hanqing Zhao et al. Multi-attentional deepfake detection. In *CVPR*, pages 2185–2194, 2021.
19. Huy H Nguyen et al. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP*, pages 2307–2311. IEEE, 2019.