Expert Systems

WILEY

Expert Systems

REVIEW

Explainable Artificial Intelligence for Deepfake Detection: Pipeline, Open Source and Comparisons

Hao Li[1,2] | Junxin Chen[1] | Bo Wang[3] | Salvador García[4] | David Camacho[5]

[1]School of Software, Dalian University of Technology, Dalian, China | [2]School of Computing Science, University of Glasgow, Glasgow, UK | [3]School of Information and Communication Engineering, Dalian University of Technology, Dalian, China | [4]Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), Granada, Spain | [5]School of Computer Systems Engineering, Universidad Politecnica de Madrid, Madrid, Spain

Correspondence: Junxin Chen (junxinchen@ieee.org)

Received: 16 October 2025 | **Revised:** 20 January 2026 | **Accepted:** 24 January 2026

Keywords: deepfake detection | explainable artificial intelligence | interpretable machine learning | quantitative evaluation methods

ABSTRACT

Deepfake detection models achieve high accuracy, yet their interpretability remains underexplored. This study presents a unified evaluation pipeline for post hoc visual explanations, grounded in the Co-12 attributes of explanation quality and operationalised through a structured framework that assesses Coherence, Composition, Correctness, Completeness, Compactness, Covariate completeness and Continuity. Applying this pipeline to 16 representative explanation methods reveals systematic differences across methodological categories. CAM-based approaches demonstrate strong spatial coherence and temporal continuity; gradient-based techniques such as Guided Backprop and LRP yield compact and accurate attributions; and redistribution-based methods including ExcitationBP and Deep Taylor maintain high consistency across evaluation conditions. In contrast, perturbation-based approaches such as SHAP and LIME exhibit weaker localisation and reduced temporal stability. By enabling controlled, attribute-level comparison of explanation methods, the proposed pipeline bridges conceptual interpretability frameworks and empirical analysis, offering practical guidance for the development and deployment of interpretable deepfake detectors in forensic and auditing applications. The source code is publicly available at: https://github.com/junxinchenieee/EAI-Deepfake-Detection.

1 | Introduction

In the absence of robust and trustworthy detection mechanisms and regulatory safeguards, the rapid dissemination of synthetic media across entertainment, social platforms and online communication poses substantial risks to public trust, political stability and personal privacy (Liz-López et al. 2024; Habbal et al. 2024; Chen et al. 2023). To mitigate such risks, watermarking and provenance techniques have been proposed to support copyright protection and content traceability—for example, uncertainty-aware watermarking for 3D Gaussian Splatting (Huang et al. 2024). Because many maliciously generated media are created without embedded identifiers, however, these techniques cannot provide complete coverage in practice, making detection indispensable for accountability and forensic reliability.

Deepfake detection has therefore become a highly active research field (Heidari et al. 2024), with deep learning–based detectors constituting the predominant approach (accounting for around 77% of published studies from 2018 to 2020 (Rana et al. 2022)) and achieving strong performance on benchmarks such as FaceForensics++ (FF++) (Rössler et al. 2019), Celeb-DF (Li, Yang, et al. 2020) and the Deepfake Detection Challenge (DFDC) dataset (Dolhansky et al. 2020; Passos et al. 2024). However, current evaluation practices rely predominantly on coarse-grained classification metrics such as accuracy or Area Under the Receiver Operating Characteristic Curve (AUC) (Barredo Arrieta et al. 2020; Yoon

et al. 2024), under which existing detectors cannot disclose the evidential basis of their decisions, lack spatial localisation of manipulated regions and do not capture temporal or distributional consistency across frames and sources (Nauta et al. 2023; Angelov et al. 2021; Linardatos et al. 2021).

While such metrics offer a coarse estimate of predictive capability, they remain insufficient in high-stakes applications—including media forensics, regulatory auditing and compliance—where decision reliability is paramount (Habbal et al. 2024; Rosso et al. 2024; Camilleri 2024). This limitation is further evident in safety-critical domains such as medical imaging (Qureshi et al. 2025; Wani et al. 2024), autonomous systems and other contexts of deep learning deployment, where erroneous predictions can have ethical, legal, or safety consequences (Chen et al. 2025; Júnior et al. 2024; Singh et al. 2025). These considerations motivate the development of detection systems that are both robust and explainable.

Explainable Artificial Intelligence (XAI) has been developed to address these shortcomings by providing insights into model decisions (Adadi and Berrada 2018; Fang et al. 2024; Cortés et al. 2024). Post hoc explanation methods include gradient- and relevance-based propagation, class activation mapping (CAM), perturbation-based methods and example-based approaches (Belaid et al. 2023; Coroama and Groza 2022). These methods differ in assumptions, computational cost and output format and their misuse can lead to misleading interpretations. Evaluation practice is fragmented: many studies rely on anecdotal visualisations or small-scale user studies, and there is still no consensus on standardised quantitative metrics (Barredo Arrieta et al. 2020; Nauta et al. 2023; Al-Ansari et al. 2024; Lage et al. 2019). Although prior work has identified attributes such as completeness, correctness, compactness and robustness, agreement on evaluation protocols remains limited (Barredo Arrieta et al. 2020; Girón et al. 2025). Recent efforts have called for reproducible, multi-metric evaluation frameworks to assess explanation reliability and make results more comparable (Nauta et al. 2023; Li, Liu, et al. 2020).

In the deepfake detection setting, explainability has particular value. It enables identification of spurious cues and dataset biases (Nauta et al. 2023; Linardatos et al. 2021), provides spatial localisation to support forensic evidence (Angelov et al. 2021; Xu et al. 2019) and allows human reviewers to verify model predictions. It also reduces compliance risks by supplying auditable decision bases (Mansoor and Iliev 2025; Malolan et al. 2020). However, without a unified and reproducible evaluation standard, it is difficult to compare explanation methods or establish their credibility and deployability in practice.

This article presents a systematic evaluation of post hoc visual explanation methods for deepfake detection. Building on the DeepfakeBench benchmark framework (DeepfakeBench) (Yan et al. 2023), which provides standardised data processing and baseline detectors, we implement 16 representative explainers on the Xception detector tested on FaceForensics++ Deepfakes and Celeb-DF-v2. We design an evaluation pipeline including four components aligned with the Co-12 criteria (a 12-property framework for evaluating explanation quality) (Nauta et al. 2023) to quantify seven dimensions of explanation quality, including correctness, completeness, compactness, covariate completeness, continuity, composition and coherence. The resulting evaluation suite offers a unified and reproducible pipeline and is readily extensible to additional detectors and datasets supported by DeepfakeBench (Yan et al. 2023).

The contributions of this article are threefold:

- **Unified evaluation pipeline**. We construct a unified evaluation pipeline for post hoc visual explanations under fixed detector (Xception (Chollet 2017)) and datasets (FF-DF (Rössler et al. 2019) and Celeb-DF-v2 (Li, Yang, et al. 2020)) conditions. All methods are executed to produce standardised pixel-level heatmaps with unified spatial resolution, normalisation and masking operations. This harmonisation reduces implementation-level variability and improves comparability across algorithms.
- **Systematic comparison across methodologies**. The evaluation includes three representative categories of post hoc explainability methods: (1) gradient- and relevance-propagation methods, (2) class activation mapping and aggregation methods and (3) perturbation-based attribution methods. A unified API and batch-mode inference pipeline ensure consistent execution across all methods under identical detection settings.
- **Implementation of quantitative evaluation framework**. Co-12 defines 12 key properties for evaluating the quality of post hoc explanations (Nauta et al. 2023). The Co-12 framework is instantiated through four measurable protocols tailored to the deepfake detection setting. Each quality criterion is formalised with a computable metric and integrated into reproducible evaluation scripts. The resulting benchmark supports fairer, replicable and extensible evaluation across diverse detectors and datasets in DeepfakeBench (Yan et al. 2023).

The remainder of this article is structured as follows. Section 2 introduces the explainable deepfake detection model. It gives background on XAI methods in this area and describes both the detector backbone and the 16 explanation methods we implement. Section 3 presents the evaluation framework. It starts with background material, then outlines the dataset and the pipeline, which puts seven evaluation attributes into practice through four components. Section 4 provides a comparison of the 16 explanation methods against the evaluation attributes, including detailed results and discussion. Finally, Section 5 summarises the main contributions, noting limitations and suggesting directions for future work. The algorithm details are available in Appendix A.

2 | Explainable Deepfake Detection Model

2.1 | Related Works

2.1.1 | Deepfake Detection Methods

Deep learning approaches are generally grouped into three categories. *Naive detectors* perform end-to-end training on facial frames using convolutional neural network backbones (CNNs), such as MesoNet, Xception or EfficientNet (Pan et al. 2020;

Sudarshana and Vamsidhar 2025). On FaceForensics++ (Yan et al. 2023), Xception reports mean AUC values of about 0.94 within the domain but drops to around 0.77 on cross-domain datasets such as Celeb-DF and DFDC, indicating sensitivity to domain shift (Yan et al. 2023). *Spatial-domain detectors* explicitly model manipulated regions or structures, for example, via capsule networks (CapsNets) (Nguyen et al. 2019), boundary detection (Chen et al. 2022), or attention mechanisms. These approaches improve localisation but remain affected by compression and unseen manipulations (Yan et al. 2023; Pan et al. 2020). Frequency-based representations have been widely applied in diverse machine learning tasks due to their ability to reveal structured patterns and invariant cues (Chen et al. 2024). As many deepfakes introduce artefacts across multiple modalities, including visual boundary inconsistencies, audio formant distortions and cross-modal rhythm mismatches, frequency-domain analysis has emerged as an effective approach for uncovering hidden manipulation patterns. *Frequency-domain detectors* leverage spectral characteristics to identify generative traces (Yan et al. 2023), employing methods such as Spatial Pyramid Structure Learning (SPSL) (Liu et al. 2021), Steganalysis Rich Model (SRM) (Luo et al. 2021), or learnable filters. These techniques often demonstrate robustness to compression and resampling; however, their performance may still be influenced by variations in recording devices and acquisition pipelines (Gao et al. 2024; Xu et al. 2025).

In addition, recent studies pursue deepfake detection from a cross-modal consistency perspective, leveraging semantic and temporal constraints across vision, speech and text modalities to improve cross-domain robustness and interpretability (Li et al. 2024; Yoon et al. 2024; Keita et al. 2025).

Traditional machine learning approaches rely on handcrafted descriptors combined with shallow classifiers (Heidari et al. 2024; Yazdinejad et al. 2020). Features include facial geometry, head pose, eye blinking, lip motion (Matern et al. 2019; Yang et al. 2018; Fang et al. 2024), texture descriptors (Guarnera et al. 2020a) and frequency statistics (Rana et al. 2022; Bonomi et al. 2020). These methods are computationally efficient and yield transparent decision chains, but their accuracy strongly depends on feature design and they generalise poorly to unseen manipulation types (Rana et al. 2022).

Video-based methods provide complementary evidence by modelling temporal consistency and motion patterns (Tariq et al. 2020; Masi et al. 2020). Approaches integrate convolutional features with recurrent networks, apply three-dimensional convolutions (3D CNNs), or analyse optical flow and motion discrepancies (Heidari et al. 2024; Sandhya and Kashyap 2025; Ciftci et al. 2024). Physiological cues such as eye-blinking periodicity and audio–visual lip synchronisation have also shown discriminative value (Jafar et al. 2020; Zhao et al. 2021). These methods capture temporal dependencies but may be affected by frame-rate changes, editing and compression noise (Rana et al. 2022; Chintha et al. 2020; Sabir et al. 2019).

Statistical forensics targets low-level signal traces that are less dependent on semantic content (Rana et al. 2022). Typical methods analyse sensor fingerprints such as Photo Response Non-Uniformity (PRNU) (Anusha and Srinagesh 2025) or distributional distances such as Kullback–Leibler (KL) divergence and Jensen–Shannon (JS) divergence (Heidari et al. 2024; Rana et al. 2022). Some approaches extract convolutional traces or fingerprints with Expectation–Maximisation (EM) to distinguish generative models (Rana et al. 2022; Guarnera et al. 2020b). While relatively stable across content types, these methods are sensitive to compression and device variation and often require preprocessing or calibration.

Beyond supervised classifiers, recent work exploits predictive uncertainty as a signal of detection under distribution shift, motivated by the intrinsic discrepancy between natural images and AI-generated images. Images inducing high uncertainty in large-scale vision models are identified as AI-generated, yielding a simple yet effective detector across benchmarks (Nie et al. 2024).

In general, deep learning methods report higher accuracy and AUC under controlled conditions than traditional or statistical approaches (Gao et al. 2024; Guo et al. 2024). However, performance drops markedly in cross-domain evaluations, reflecting limited robustness to domain shifts (Rana et al. 2022). Reported results also vary due to differences in datasets, compression levels and evaluation metrics (Yan et al. 2023; Xu et al. 2024). A unified data management pipeline, training strategy and evaluation protocol are therefore needed to ensure reproducible and generalisable conclusions for real-world deployment.

2.1.2 | XAI Methods: Taxonomy and Mainstream

Explainable AI (XAI) methods can be organised according to several complementary dimensions. First, in terms of *timing*, they are commonly distinguished into *ante-hoc*, *in-training* and *post hoc* methods (Barredo Arrieta et al. 2020; Dwivedi et al. 2023). Second, *scope* distinguishes between global methods, which approximate the overall decision structure and feature–output relations and local methods, which focus on explanations for individual predictions or small neighbourhoods (Chen, Lundberg, and Lee 2019). Third, *coupling* differentiates between model-specific approaches that rely on internal signals (e.g., gradients, propagation rules) and model-agnostic approaches that operate under a black-box assumption and use only input–output behaviour (Simonyan et al. 2013; Springenberg et al. 2014). Fourth, *form* refers to the representation of the explanation, including saliency maps (Chen, Lundberg, and Lee 2019), surrogate models, examples or prototypes, feature relevance scores (Chattopadhay et al. 2018), or textual rationales (Barredo Arrieta et al. 2020). Finally, transparency can be considered at the level of entire models (intrinsically interpretable), subsystems (decomposability) or post hoc analysis of black-box predictors (Holzinger et al. 2022).

Ante-hoc methods construct inherently interpretable models whose internal mechanisms are transparent by design and therefore directly inspectable by humans (Barredo Arrieta et al. 2020). Typical representatives include linear and logistic regression, decision trees, rule-based systems, generalised additive models (GAMs) (Caruana et al. 2015) and Bayesian networks (Barredo Arrieta et al. 2020). In linear and logistic regression, the prediction is expressed as a weighted sum of the input features. The

model coefficients can therefore be interpreted as estimates of each feature's contribution, under suitable regularity conditions. Decision trees encode decision logic as human-readable paths from the root to the leaves. Generalised additive models decompose the prediction into a sum of low-dimensional shape functions that can be visualised and scrutinised by domain experts. Bayesian networks, in turn, provide a probabilistic graphical representation of causal and conditional dependencies, thereby supporting transparent reasoning under uncertainty. Beyond these classical models, a body of work has sought to retain white-box interpretability while improving predictive performance. Supersparse Linear Integer Models (SLIM), which constrain predictors to simple integer-weighted scoring systems to maximise readability for domain users (Ustun and Rudin 2016). GA2Ms that extend GAMs with automatically learned pairwise interactions trained using modern boosting strategies (Lou et al. 2013). Boolean Rule Column Generation frameworks that learn compact disjunctive normal form (DNF) or conjunctive normal form (CNF) rule sets balancing rule simplicity and predictive fit (Dash et al. 2018). Generalised Linear Rule Models, which combine Generalised Linear Models (GLMs) estimation with rule-based feature construction to capture nonlinear structure while remaining human-interpretable (McCullagh 2019). Related advances in decision-tree design—such as incorporating monotonicity constraints and label-dependency fusion for multilabel tasks—further aim to strengthen both interpretability and predictive performance (Khalaf et al. 2026). Nevertheless, intrinsically interpretable models typically sacrifice representational capacity and accuracy in high-dimensional vision tasks where complex feature hierarchies and interactions dominate (Barredo Arrieta et al. 2020; Gunning and Aha 2019). As a result, deep neural networks remain the dominant detectors in computer vision and media forensics. However, for forensic and regulatory applications, the algorithmic transparency, auditability, stable global logic and ease of communicating feature or rule contributions make these inherently interpretable models particularly valuable where justification, contestability and accountability are mandated (The Royal Society 2019).

In-training explainability methods introduce interpretability objectives directly into the optimisation process. Some approaches impose sparsity-inducing or disentanglement-oriented regularisation on internal representations, encouraging latent factors and features to align with distinct, human-interpretable structures (Higgins et al. 2017; Chen et al. 2016). Concept Bottleneck Models learn an explicit intermediate representation consisting of human-defined concepts and perform prediction solely on this concept space (Koh et al. 2020). Prototype-based networks integrate a prototype layer during learning, so that class decisions are formed through similarity to learned prototypical parts (Chen, Li, et al. 2019). Attention-aligned training constrains gradient- or attention-based importance maps to match human-annotated regions during optimisation (Ross et al. 2017). Modular and factorised architectures further enforce interpretable structure through functional decomposition during training (Andreas et al. 2016). Other approaches introduce explanation-based regularisers that penalise reliance on features identified as irrelevant by human supervision, or embed explicit concept-weight parameterisations whose stability and fidelity are enforced by additional loss terms (Ross et al. 2017; Alvarez-Melis and Jaakkola 2018). Across these methods, interpretability constraints are embedded into the training objective or architecture design, rather than being applied post hoc.

Post hoc methods generate explanations after training and are widely applied in computer vision, as state-of-the-art detectors are typically complex neural networks that are not inherently transparent. Within this setting, post hoc local explanation methods dominate. Model-specific post hoc methods derive explanations by exploiting internal network signals such as gradients, activations, or intermediate feature representations. Gradient- and propagation-based approaches estimate the sensitivity of outputs to input perturbations, producing saliency maps that highlight class-discriminative regions (Angelov et al. 2021). Variants such as Integrated Gradients (Sundararajan et al. 2017), DeepLIFT (Shrikumar et al. 2017), Layer-wise Relevance Propagation (Bach et al. 2015) and Deep Taylor Decomposition (Kim et al. 2018) introduce principled propagation rules or axiomatic guarantees to stabilise attribution (Linardatos et al. 2021). Extensions including SmoothGrad (Smilkov et al. 2017) and PatternNet (Kindermans et al. 2017) reduce noise and correct estimator bias. More recently, Transformer-oriented extensions such as GradToken64 adopt class-aware gradient decoupling to disentangle the semantics of the class token and produce category-specific relevance maps, thereby improving localisation accuracy and reliability compared to CNN-based counterparts (Cheng et al. 2025). The class activation category (CAM, Grad-CAM, Grad-CAM++) projects class information onto convolutional features to produce heatmaps. These methods are efficient but limited in spatial precision due to upsampling artefacts (Linardatos et al. 2021). In parallel, concept-based attribution methods seek to explain predictions in terms of human-understandable semantic concepts encoded in intermediate feature spaces rather than individual pixels. Testing with Concept Activation Vectors (TCAV) quantifies the directional sensitivity of predictions to concept vectors derived from exemplar sets, while Automated Concept Extraction (ACE) extends this paradigm by automatically discovering coherent visual concepts before applying TCAV (Kim et al. 2017). Model-agnostic perturbation methods provide an alternative by directly probing input–output behaviour. Randomised Input Sampling for Explanation (RISE) (Petsiuk et al. 2018) aggregates predictions over random masks to obtain high-fidelity saliency maps at a higher computational cost. Local Interpretable Model-agnostic Explanations (LIME) explains single predictions through local perturbations and linear surrogates, while SHapley Additive exPlanations (SHAP) unifies additive attribution with game-theoretic guarantees (local accuracy, missingness, consistency). Both approaches are widely applied in black-box analysis, though they exhibit sensitivity to parameter choices and assumptions about feature independence. Beyond these mainstream methods, a broad range of explanation paradigms has been proposed, including Anchors (Ribeiro et al. 2018) for high-precision rules, contrastive and counterfactual explanations (Dhurandhar et al. 2018), prototype and criticism selection (Gurumoorthy et al. 2019), permutation importance (Altmann et al. 2010), subset-selection methods such as Learning to eXplain (L2X) (Chen et al. 2018) and effect-visualisation techniques such as Partial Dependence Plots (PDP), Individual Conditional Expectation (ICE) and Accumulated Local Effects (ALE) (Friedman 2001; Goldstein et al. 2015). Some recent studies have also explored generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) (Charte et al. 2018) for explaining data-driven decisions.

Table 1 consolidates this taxonomy, highlighting scope, coupling, modality and typical applicability in the present project.

2.1.3 | Applications in Deepfake Detection

Most studies apply post hoc visual explanations to deepfake detectors such as Xception, EfficientNet and Vision Transformers (Malolan et al. 2020; Gowrisankar and Thing 2024). Explanations are typically presented as heatmaps or local masks that highlight manipulated regions (Aghasanli et al. 2023). Early work emphasised interpretability and robustness over benchmark accuracy, using methods such as LIME, LRP and Grad-CAM, and testing stability under Gaussian blur and affine transformations (Haq et al. 2024; Jayakumar and Skandhakumar 2022). A recurring observation is that explanations often highlight facial areas such as the nose and mouth where artefacts are common (Malolan et al. 2020). Later studies extended this approach with EfficientNet and Xception backbones, using Grad-CAM, Anchors and LIME, but many evaluations remained qualitative and limited to small samples (Silva et al. 2022).

To reduce evaluation bias, task-aligned quantitative frameworks have been proposed. Direct deletion or blurring can distort faces and yield misleading results, so one approach applies adversarial perturbations with NES inside salient regions and measures confidence drops as evidence effectiveness. Using this method, Grad-CAM++, RISE, SHAP, LIME and Sobol' sensitivity analysis were compared, with LIME showing competitive results across multiple manipulation types (Tsigos et al. 2024). Another line of work uses Network Dissection to link internal units to named facial concepts through intersection over union, providing semantic context in addition to raw heatmaps (Mansoor and Iliev 2025). Studies on StyleGAN-based datasets also applied LIME to confirm local evidence supporting high classification accuracy across convolutional networks, improving credibility (Khan 2023).

In general, XAI research on deepfake detection has progressed from qualitative visualisation to quantitative evaluation. Remaining challenges include reliance on small qualitative samples, limited assessment of temporal consistency in videos and inconsistent evaluation protocols and cost reporting. These gaps highlight the need for unified comparisons under controlled conditions, with standard datasets, metrics and explicit reporting of computational cost (Mansoor and Iliev 2025; Tsigos et al. 2024; Khan 2023).

2.2 | Detector Backbone

As reviewed in Section 2.1.1, deepfake detection models are commonly built upon four major classes of backbone architectures: (i) convolutional neural network (CNN) classifiers that operate on individual facial frames, (ii) spatial-domain structured models, (iii) frequency-domain CNN backbones that incorporate spectral priors and (iv) temporal or video-based backbones that exploit motion cues using 3D convolutions or recurrent networks. A large proportion of existing works employ conventional CNN classifiers such as MesoNet, Xception, or EfficientNet (Pan et al. 2020; Sudarshana and Vamsidhar 2025), which learn semantic representations directly from facial frames. Spatial-domain detectors instead model manipulated regions or structures via capsule networks, boundary-aware learning or attention mechanisms (Nguyen et al. 2019; Chen et al. 2022). Frequency-domain detectors leverage spectral characteristics and handcrafted or learnable filters such as SPSL and SRM to capture generative traces (Liu et al. 2021; Luo et al. 2021; Gao et al. 2024; Xu et al. 2025). Temporal and video-based methods further incorporate motion dynamics using recurrent networks, 3D CNNs, or optical-flow-based architectures (Tariq et al. 2020; Masi et al. 2020; Sandhya and Kashyap 2025; Ciftci et al. 2024). Among these categories, CNN-based detectors remain the most extensively adopted and empirically validated family in current deepfake detection research.

Within this category, Xception belongs to the group of lightweight CNN architectures constructed from depthwise separable convolutions. Architectures of similar design, including EfficientNet and modern ResNet variants, have also been widely applied as backbones in deepfake detection. In this work, Xception is adopted as the CNN backbone for the explainable detection model due to its widespread use in deepfake detection research and its consistently strong performance across standard benchmarks, which makes it a suitable and representative baseline (Yan et al. 2023). Employing a single, representative CNN backbone also prevents detector variability from acting as a confounding factor, allowing the explainability methods to be compared within a controlled yet practically relevant detection setting. Xception serves as the default detector in our work and the proposed explainability pipeline can also be applied to other backbone networks. Accordingly, this study adopts the Xception-based binary detector within the DeepfakeBench framework.

The detector is kept frozen at inference time with publicly released best-performing weights. Inputs are aligned face frames, where each video contributes a fixed set of 32 frames. Each RGB frame is resized to 256×256 and normalised with a channelwise mean and standard deviation of 0.5. Masks and landmarks provided with the data are reserved for evaluation and do not participate in the forward pass. A single forward pass produces two logits and the corresponding fake class probability $p \in [0, 1]$. The high level feature map right before global average pooling is also retained for later spatial alignment and visualisation. Unless a hard decision is explicitly required, metrics and explainability computations are based on the continuous probability p or on the corresponding logit, with a default threshold of 0.5 used only for hard decisions.

The network follows the standard Xception design, built around depthwise separable convolutions with residual connections (Chollet 2017). The architecture can be divided into entry, middle and exit stages. The entry stage performs initial downsampling with a strided convolution and expands channels. The middle stage stacks residual blocks with a constant channel width to deepen the receptive field while largely preserving spatial resolution. The exit stage applies another downsampling step, followed by two separable convolutions that lift the channel dimension to 1536 and 2048, then uses global average pooling

TABLE 1 | Consolidated taxonomy of explainability methods across timing, coupling, modality and scope.

Method	Timing	Coupling	Image	Tabular	Text	Local	Global	In project
Linear/logistic regression	Intrinsic	NA	✗	✓	✓	✓	✓	✗
Decision trees	Intrinsic	NA	✗	✓	✗	✓	✓	✗
Rule-based models	Intrinsic	NA	✗	✓	✗	✓	✓	✗
GAMs	Intrinsic	NA	✗	✓	✗	✓	✓	✗
KNN	Intrinsic	NA	✓	✓	✓	✓	✗	✗
Bayesian networks	Intrinsic	NA	✗	✓	✗	✓	✓	✗
SLIM	Intrinsic	NA	✗	✓	✓	✓	✓	✗
Boolean rule column generation	Intrinsic	NA	✗	✓	✗	✓	✓	✗
Rule ensembles	Intrinsic	NA	✗	✓	✗	✓	✓	✗
β-VAE	In-training	Model-specific	✓	✗	✗	✗	✓	✗
InfoGAN	In-training	Model-specific	✓	✗	✗	✗	✓	✗
Concept bottleneck models	In-training	Model-specific	✓	✓	✗	✓	✓	✗
ProtoPNet	In-training	Model-specific	✓	✗	✗	✓	✗	✗
Attention-aligned training	In-training	Model-specific	✓	✓	✓	✓	✗	✗
Neural module networks	In-training	Model-specific	✓	✗	✗	✓	✓	✗
Explanation-regularised training	In-training	Model-specific	✓	✓	✓	✓	✗	✗
Vanilla gradient	Post hoc	Model-specific	✓	✗	✗	✓	✗	✓
Integrated Gradients	Post hoc	Model-specific	✓	✓	✓	✓	✗	✓
SmoothGrad	Post hoc	Model-specific	✓	✗	✗	✓	✗	✓
Input ✗ Gradient	Post hoc	Model-specific	✓	✓	✓	✓	✗	✓
Guided Backprop	Post hoc	Model-specific	✓	✗	✗	✓	✗	✓
Deconvolution	Post hoc	Model-specific	✓	✗	✗	✓	✗	✓
LRP	Post hoc	Model-specific	✓	✗	✗	✓	✗	✓
Deep Taylor Decomposition	Post hoc	Model-specific	✓	✗	✗	✓	✗	✓
ExcitationBP	Post hoc	Model-specific	✓	✗	✗	✓	✗	✓
DeepLIFT	Post hoc	Model-specific	✓	✓	✓	✓	✗	✗
PatternNet	Post hoc	Model-specific	✓	✗	✗	✓	✗	✗
PatternAttribution	Post hoc	Model-specific	✓	✗	✗	✓	✗	✗
CAM	Post hoc	Model-specific	✓	✗	✗	✓	✗	✓
Grad-CAM	Post hoc	Model-specific	✓	✗	✗	✓	✗	✓
Grad-CAM++	Post hoc	Model-specific	✓	✗	✗	✓	✗	✓
Score-CAM	Post hoc	Model-specific	✓	✗	✗	✓	✗	✓
GradToken64	Post hoc	Model-specific	✓	✗	✗	✓	✗	✓
TCAV	Post hoc	Model-specific	✓	✓	✗	✓	✓	✗
ACE	Post hoc	Model-specific	✓	✓	✗	✓	✓	✗
Occlusion Sensitivity	Post hoc	Model-agnostic	✓	✗	✗	✓	✗	✓

(Continues)

TABLE 1 | (Continued)

Method	Timing	Coupling	Image	Tabular	Text	Local	Global	In project
RISE	Post hoc	Model-agnostic	✓	✗	✗	✓	✗	✗
LIME	Post hoc	Model-agnostic	✓	✓	✓	✓	✗	✓
DLIME	Post hoc	Model-agnostic	✓	✓	✓	✓	✗	✗
SHAP	Post hoc	Model-agnostic	✓	✓	✓	✓	✓	✓
SHAP TreeSHAP	Post hoc	Model-specific	✗	✓	✗	✓	✓	✗
Anchors	Post hoc	Model-agnostic	✓	✓	✓	✓	✗	✗
CEM	Post hoc	Model-agnostic	✓	✓	✗	✓	✗	✗
Counterfactual Explanations	Post hoc	Model-agnostic	✓	✓	✓	✓	✗	✗
Prototype & Criticism	Post hoc	Model-agnostic	✓	✓	✗	✓	✓	✗
Permutation Importance	Post hoc	Model-agnostic	✗	✓	✗	✗	✓	✗
L2X	Post hoc	Model-agnostic	✓	✓	✗	✓	✗	✗
PDP	Post hoc	Model-agnostic	✗	✓	✗	✗	✓	✗
ICE	Post hoc	Model-agnostic	✗	✓	✗	✓	✗	✗
ALE	Post hoc	Model-agnostic	✗	✓	✗	✗	✓	✗
LIVE	Post hoc	Model-agnostic	✗	✓	✗	✓	✗	✗
breakDown	Post hoc	Model-agnostic	✗	✓	✗	✓	✗	✗
ProfWeight	Post hoc	Model-agnostic	✓	✓	✓	✗	✓	✗

and a linear classifier to yield two logits. The feature map immediately before the classifier has a spatial size of approximately 8×8 with 2048 channels, which is suitable for class activation mapping and related localisation methods (Yan et al. 2023). The activation function is ReLU.

In this work, the Xception detector is trained on the FF++ dataset. On the DeepFakes subset of FF++ (FF-DF), frame-level performance computed from continuous probabilities is as follows: accuracy 0.877, area under the curve 0.988, equal error rate 0.056 and average precision 0.989. The video-level area under the curve is 0.997. To assess cross-dataset generalisation, the same pretrained model is evaluated on Celeb-DF-v2, where the frame-level accuracy, AUC, equal error rate and average precision are 0.695, 0.740, 0.329 and 0.837, respectively, and the video-level AUC is 0.816.

2.3 | Explainability Methods

2.3.1 | Overview

This work evaluates 16 explainability methods spanning three major categories:

1. **Gradient and relevance-based methods:** Vanilla Gradient, Integrated Gradients, SmoothGrad, Input × Gradient, Guided Backpropagation (Guided Backprop), Deconvolution, Layer-wise Relevance Propagation (LRP), Deep Taylor Decomposition (DeepTaylor), Excitation Backpropagation (ExcitationBP).
2. **Class activation mapping:** Classic CAM, Grad-CAM, Grad-CAM++, Score-CAM.
3. **Perturbation-based methods:** Occlusion Sensitivity, LIME, KernelSHAP.

All methods are standardised to output attribution maps in the same tensor format, ensuring comparability across categories. Furthermore, all explainers are integrated into a unified interface and method loader, enabling reproducible evaluation and consistent switching across methods.

2.3.2 | Gradient and Relevance-Based Explanation Methods

Vanilla Gradient.

This method computes input gradients with respect to the scalar output score of the fake class, denoted by $\phi(x) = p_{\text{fake}}(x)$, where $x \in \mathbb{R}^{B\times 3\times H\times W}$ represents a batch of input images. The scalar objective is defined as the sum over all sample-wise class scores,

$$J(x) = \sum_{b=1}^{B} \phi\left(x^{(b)}\right).$$

Gradients are computed by backpropagating $J(x)$ through a frozen model, resulting in a gradient tensor $G = \frac{\partial J(x)}{\partial x} \in \mathbb{R}^{B\times 3\times H\times W}$. The absolute value of this tensor is taken to obtain the saliency map $A = |G|$. Channel-wise aggregation and per-image

min–max normalisation are applied for visualisation. The related procedure is provided as Algorithm A1 in Appendix A.

Integrated Gradients.

To address issues like gradient saturation and local discontinuities, Integrated Gradients estimates feature attributions by accumulating gradients along a linear path from a baseline input x' (zero image) to the actual input x. The method computes:

$$\mathrm{IG}_i(x) = \left(x_i - x_i'\right) \int_0^1 \frac{\partial \phi\left(x' + \alpha\left(x - x'\right)\right)}{\partial x_i} d\alpha,$$

which is approximated via discrete summation over T steps (Sundararajan et al. 2017). In our experiments, we use the library default of $T = 50$. The absolute attribution values $|\,\mathrm{IG}(x)\,|$ are aggregated across channels and normalised on a per-image basis. Compared to Vanilla Gradient, Integrated Gradients yield more stable and interpretable results under input scaling.

SmoothGrad.

SmoothGrad enhances gradient-based explanations by reducing noise in the resulting saliency maps (Smilkov et al. 2017). The method generates multiple perturbed versions of the input by adding Gaussian noise $\varepsilon^{(t)} \sim \mathcal{N}\left(0, \sigma^2 I\right)$ and computes gradients for each,

$$\mathrm{SG}(x) = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{\partial \phi\left(x + \varepsilon^{(t)}\right)}{\partial x} \right|.$$

In our implementation, we use $n = 25$ samples and a noise standard deviation of $\sigma = 0.15$. The averaged attribution maps are then post-processed through channel-wise maximum aggregation and min–max normalisation.

Input × Gradient.

This method multiplies each input pixel by the corresponding partial derivative of the model output with respect to that pixel (Kokhlikyan et al. 2020):

$$A(x) = x \odot \frac{\partial \phi(x)}{\partial x}, \quad A \in \mathbb{R}^{B \times 3 \times H \times W}.$$

The result is interpreted as a feature-wise contribution score. The final attribution map $|\,A(x)\,|$ is aggregated and normalised as in previous methods. This approach highlights salient features where both input intensity and model sensitivity are high.

Guided Backprop.

Guided Backprop modifies the gradient flow during backpropagation by enforcing a double gating mechanism at ReLU activations (Linardatos et al. 2021): only units that are both positively activated and receive a positive gradient are allowed to propagate their gradients:

$$\frac{\partial \phi}{\partial a} \leftarrow \frac{\partial \phi}{\partial a} \cdot \mathbf{1}(a > 0) \cdot \mathbf{1}\left(\frac{\partial \phi}{\partial a} > 0\right).$$

The resulting gradients are taken in absolute value and processed similarly to prior methods. This mechanism produces sharper and more interpretable saliency maps that are aligned with edges and textures, though it lacks class discriminativity.

Deconvolution.

Deconvolution applies a modified backpropagation scheme where only positive gradients are allowed to pass through ReLU units, regardless of the forward activation (Linardatos et al. 2021):

$$\frac{\partial \phi}{\partial a} \leftarrow \frac{\partial \phi}{\partial a} \cdot \mathbf{1}\left(\frac{\partial \phi}{\partial a} > 0\right).$$

This results in smoother but less localised saliency maps compared to Guided Backprop. Aggregation and normalisation steps follow the standard pipeline.

LRP.

LRP redistributes the model's output backward through the network in a way that preserves the total prediction score (Bach et al. 2015; Anders et al. 2021). For each sample, we set the target class to fake and propagate relevance using the EpsilonPlus rule C:

$$A = \mathrm{ZGradient}(f, C)(x, t).$$

The resulting tensor $A \in \mathbb{R}^{B \times 3 \times H \times W}$ is upsampled to 256×256, then aggregated spatially as $S_b(u, v) = \sum_{c=1}^{3} A_{b,c}(u, v)$. Final maps $\widehat{S}_b$ are obtained via min–max normalisation with ϵ-stabilisation. LRP provides relevance maps that are more consistent and class-specific. The related procedure is provided as Algorithm A2 in Appendix A.

DeepTaylor.

This method is derived from Taylor decomposition, interpreting the model's prediction as a function to be expanded around a root point (Kim et al. 2018). In practice, we adopt the EpsilonPlus rule for stable attribution:

$$A = \mathrm{ZGradient}(f, C)(x, \phi), \quad A^{\uparrow} = \mathrm{Interp}_{256 \times 256}(A),$$
$$S_b(u, v) = \sum_c A^{\uparrow}_{b,c,u,v}, \quad \widehat{S}_b = \mathrm{MinMaxNorm}\left(S_b\right).$$

This approximation emphasises positive, cooperative features contributing to the model's decision and is particularly robust for bounded input domains. The related procedure is provided as Algorithm A3 in Appendix A.

Excitation Backpropagation.

ExcitationBP focuses only on positively contributing paths within the network, routing evidence exclusively through excitatory activations (Linardatos et al. 2021). The method

suppresses negative evidence by using the ExcitationBackprop composite rule:

$$A = \mathrm{ZGradient}(f, C)(x, \phi), \quad A^{\uparrow} = \mathrm{Interp}_{256\times256}(A),$$
$$S_b(u,v) = \sum_c A^{\uparrow}_{b,c,u,v}, \qquad \widehat{S}_b(u,v) = \frac{S_b(u,v) - \min S_b}{\max S_b - \min S_b + \varepsilon}.$$

This rule results in heatmaps that distinctly highlight regions strongly responsible for activating the predicted class. The related procedure is provided as Algorithm A4 in Appendix A.

2.3.3 | CAM-Based Explanation Methods

Classic CAM.

Classic CAM applies when the classifier consists of a global average pooling (GAP) layer followed by a linear classifier with class-specific weights $w_k^{(y)}$ (Selvaraju et al. 2017). Given activation maps $\{A_k\}_{k=1}^{K}$ from the last convolutional layer, the class activation map for class y is computed as:

$$\mathrm{CAM}_y^{(b)} = \mathrm{ReLU}\left(\sum_{k=1}^{K} w_k^{(y)} A_k^{(b)}\right).$$

In our binary setting, we use $y = \mathrm{fake}$. The resulting map is upsampled to the input resolution and min–max normalised per image before storage. The related procedure is provided as Algorithm A8 in Appendix A.

Grad-CAM.

Grad-CAM computes class activation maps by weighting the feature maps of a convolutional layer using the global average of the gradients corresponding to a target class (Chattopadhay et al. 2018). Specifically, for a given input, the channel-wise importance weights are defined as:

$$w_k = \frac{1}{hw} \sum_{i,j} g_{kij},$$

where g_{kij} is the gradient of the class score with respect to activation map A_k. The weighted combination of feature maps is then passed through a ReLU activation:

$$\mathrm{CAM} = \mathrm{ReLU}\left(\sum_k w_k A_k\right).$$

The resulting heatmap is upsampled to the input resolution and min–max normalised per image prior to storage. This method highlights spatial regions with strong positive influence on the class prediction. The related procedure is provided as Algorithm A5 in Appendix A.

Grad-CAM++.

Grad-CAM++ extends Grad-CAM by introducing pixel-wise weighting coefficients that allow more precise localisation, particularly when multiple instances of an object are present (Zhou et al. 2016). Using squared and cubed gradients ($g^2 = g \odot g$, $g^3 = g^2 \odot g$), the importance weights are computed as:

$$\alpha_{kij} = \frac{g_{kij}^2}{2g_{kij}^2 + \sum_{u,v} A_{kuv} g_{kuv}^3 + \varepsilon}, \qquad w_k = \sum_{i,j} \alpha_{kij}\, \mathrm{ReLU}(g_{kij}),$$
$$\mathrm{CAM} = \mathrm{ReLU}\left(\sum_k w_k A_k\right),$$

where a small constant ε (e.g., 10^{-8}) is used to ensure numerical stability. As with Grad-CAM, the map is upsampled and normalised before being stored. The related procedure is provided as Algorithm A6 in Appendix A.

Score-CAM.

Score-CAM replaces gradients with forward-passing activation scores, making the approach gradient-free (Linardatos et al. 2021). It first ranks the activation channels based on their mean activation, and retains the top K channels (default $K = 32$). For each selected channel and input sample, a normalised mask is created:

$$M_k^{(b)} = \mathrm{norm}\left(\uparrow \mathrm{ReLU}\left(A_k^{(b)}\right)\right).$$

These masks are used to spatially perturb the input, and class scores are obtained by performing a forward pass with each masked input:

$$w_k^{(b)} = p_{\mathrm{fake}}\left(x^{(b)} \odot M_k^{(b)}\right), \qquad \mathrm{CAM}^{(b)} = \mathrm{ReLU}\left(\sum_{k \in \mathrm{Top}\text{-}K} w_k^{(b)} M_k^{(b)}\right).$$

To manage computational cost, masked inputs are processed in small batches (default size = 16). Final saliency maps are normalised and saved. The related procedure is provided as Algorithm A7 in Appendix A.

2.3.4 | Perturbation-Based Explanation Methods

Occlusion Sensitivity.

Occlusion sensitivity estimates feature importance by systematically masking localised regions of the input and observing the effect on the model's output score (Holzinger et al. 2022). Given a target function $\phi(x) = p_{\mathrm{fake}}(x)$, the input image is perturbed using a sliding window of size (h_w, w_w) with stride (s_h, s_w). For each windowed region S, the attribution is computed as the difference in output probability between the original and perturbed inputs:

$$\Delta_S = \phi(x) - \phi\left(x_{[S \to b]}\right),$$

where $x_{[S \to b]}$ denotes the image with region S replaced by a baseline value b (either the per-image mean or zero). The attribution Δ_S is uniformly distributed across the masked region and accumulated across overlapping windows to construct a dense saliency map. Final attributions are converted to absolute values and normalised. In our experiments, we use a default window size of 32×32 pixels, stride of 16×16 and per-image mean as

baseline. The related procedure is provided as Algorithm A9 in Appendix A.

KernelSHAP Attribution.

KernelSHAP provides a model-agnostic approximation of Shapley values by treating image patches as interpretable input features (Lundberg and Lee 2017). The input is divided into non-overlapping patches (default patch size $= 32 \times 32$), each associated with a binary feature mask shared across all colour channels. A set of randomly sampled binary coalition masks is generated, and the corresponding perturbed inputs are evaluated to obtain output scores. These are used in a weighted linear regression to estimate the Shapley value φ_j for each patch:

$$\phi(x) \approx \sum_j \varphi_j z_j,$$

where z_j indicates whether patch j is included. Patch-level attributions are broadcast back to pixel resolution and stacked over three channels. Absolute values are used before per-image normalisation and saving. The baseline image can be specified via different modes: mean image, zero image, or a fixed constant value. The related procedure is provided as Algorithm A10 in Appendix A.

LIME.

LIME approximates the model's behaviour locally by fitting a simple interpretable model (e.g., linear regression) around the input of interest (Ribeiro et al. 2016). The image is first segmented into superpixels, then randomly perturbed by turning off various combinations of segments. Each perturbed input is passed through the detector to obtain a class probability vector $[p_{\text{real}}, p_{\text{fake}}]$, where p_{fake} is used as the target. Using $N = 400$ perturbations and a batch size of 64, LIME fits a weighted linear model to approximate the local decision boundary. The resulting weights per superpixel are mapped to a low-resolution saliency map, which is then bilinearly upsampled to match the input resolution (H, W). The final map is normalised to $[0, 1]$ and replicated across colour channels for visualisation. The related procedure is provided as Algorithm A11 in Appendix A.

3 | Evaluation Framework and Measures

3.1 | XAI Evaluation

Evaluation of explainable artificial intelligence (XAI) methods typically focuses on two dimensions: distinguishing between plausibility and faithfulness, and differentiating evaluation with or without user involvement (Dwivedi et al. 2023; Al-Ansari et al. 2024). These aspects are recurrent in survey studies and provide the basis for later systematisation.

Adebayo et al. show that anecdotal visual examples, such as saliency maps, may be misleading (Adebayo et al. 2018). Leavitt and Morcos warn against equating intuitiveness with faithfulness, and Jacovi and Goldberg argue that human-perceived reasonableness does not guarantee alignment with model reasoning (Leavitt and Morcos 2020; Jacovi and Goldberg 2020). Samek et al. emphasise that explanation quality depends on both the classifier and data, while Robnik-Šikonja and Bohanec stress the independence of predictive accuracy and explanation correctness (Nauta et al. 2023; Robnik-Šikonja and Bohanec 2018; Schwab and Karlen 2019).

Doshi-Velez and Kim propose three settings—application-grounded, human-grounded and functionally grounded—and recommend systematic use of the latter when human evidence exists (Doshi-Velez and Kim 2018). Herman and Ancona et al. highlight that user studies may favour intuitive but less faithful explanations, whereas functional metrics allow falsifiable comparisons (Oztireli et al. 2019). From an HCI perspective, Greenberg and Buxton caution that premature user testing may suppress methodological innovation (Nauta et al. 2023; Greenberg and Buxton 2008).

Bibliometric surveys indicate that 33% of studies rely only on anecdotal evidence, 58% apply quantitative metrics and 22% conduct user studies, with just 23% being application-grounded. This shows that systematic quantification remains limited (Nauta et al. 2023). To address this, OpenXAI provides a benchmark integrating datasets, explanation methods and metrics across three dimensions—faithfulness, stability and fairness. It introduces the SynthGauss generator for ground-truth evaluation and supports perturbation- and subgroup-based assessments where no ground truth exists (Agarwal et al. 2022).

OpenXAI also unifies LIME, SHAP and gradient-based explainers under a single pipeline to enhance comparability (Agarwal et al. 2022). Other systematisations emphasise user-focused indicators: Tintarev et al. identify transparency, scrutability, trust, efficiency and satisfaction; Vilone et al. classify evaluation by stage, scope, modality and output; and Zhou et al. decompose evaluation into model-, attribution- and example-based approaches (Coroama and Groza 2022). Toolkits such as AIX360, XAI-Bench, DiCE and Xplique further reduce engineering barriers (Nauta et al. 2023; Dwivedi et al. 2023).

The Co-12 framework structures explanation quality into content, presentation and user dimensions, mapping metrics to properties such as correctness, completeness, consistency, continuity, contrastivity and covariate completeness (Nauta et al. 2023). Related discussions include Miller and Lakkaraju on contrastivity and Atanasova et al. on consistency; presentation and user aspects—such as compactness, coherence and controllability—are supported by studies on comprehensibility and interactive feedback (Nauta et al. 2023).

Building upon the gaps identified in the evaluation of explainable deepfake detection, this study implements a systematic evaluation of mainstream post hoc visual explanation methods within the context of deepfake detection. We employ the Xception architecture as the detector, due to its widespread use and robustness in prior work; further details are provided in Section 2.2. A total of 16 explanation methods spanning gradient-based, CAM-based and perturbation-based paradigms are included in the evaluation, as described in Section 2.3. This study is conducted on the DeepFakes subset of FF++ (FF-DF) with C23 compression and the Celeb-DF-v2 dataset, selected for its large-scale manipulated video content and compression-aware realism. Dataset details can be found in Section 3.2. This evaluation adopts the general structure of the CO-12 framework, which provides a principled basis for organising explanation properties. The specific set of properties selected

in this study, along with the corresponding evaluation protocols and implementation pipeline, are detailed in Section 3.3.

3.2 | Dataset

A wide range of public datasets have been released to support research on deepfake detection, covering diverse manipulation techniques, data scales and acquisition conditions. A small number of benchmark datasets account for the majority of experimental evaluations in this field. In particular, FF++ is the most frequently used dataset, appearing in nearly 30% of published studies, followed by Celeb-DF (~13%) and the DeepFake Detection Challenge (DFDC) dataset (about 11%), while other datasets individually account for less than 10% of usage each. Early datasets such as UADFV (Yang et al. 2019) and DeepfakeTIMIT (Korshunov and Marcel 2018) provide relatively small collections of manipulated videos generated under controlled settings and have primarily been used for proof-of-concept studies and early algorithm validation. Subsequent benchmarks substantially expanded both scale and methodological diversity, including FaceForensics (Rössler et al. 2018) and its extended version FF++ (Rössler et al. 2019), Google's DeepFake Detection (DFD) dataset (Ai 2019) and the DFDC dataset (Dolhansky et al. 2020). In parallel, a number of image-based datasets, such as SwapMe and FaceSwap (SMFW) (Zhou et al. 2017), Fake Faces in the Wild (FFIW) (Khodabakhsh et al. 2018) and Swapped Face Detection (SFD) (Ding et al. 2019), were introduced to study facial manipulation detection at the image level. More recent datasets further emphasise realism, diversity, or forgery coverage, including Celeb-DF and its refined version Celeb-DF-v2 (Li, Gao, et al. 2020; Li, Yang, et al. 2020), DeeperForensics-1.0 (Jiang et al. 2020), WildDeepfake (Zi et al. 2020) and the recently proposed DF40 benchmark (Yan et al. 2024), which collectively reflect a broad spectrum of synthesis quality, identity diversity, recording conditions and manipulation techniques.

These datasets differ substantially in their design objectives and annotation granularity. FF++ is constructed from 1000 real videos and multiple face manipulation methods, and provides pixel-level masks of manipulated regions together with multiple compression levels (Rössler et al. 2019). This structured design facilitates controlled evaluation of detection performance and supports localisation-based and explainability-oriented analysis. Celeb-DF-v2 is built from real-world celebrity videos collected from online sources and includes matched pairs of pristine and manipulated videos generated using improved synthesis pipelines (Li, Yan, et al. 2020). Compared with FF++, Celeb-DF-v2 exhibits higher visual fidelity and fewer obvious artefacts, making it a commonly adopted benchmark for evaluating cross-dataset generalisation. DFDC, in contrast, emphasises large-scale diversity, with thousands of videos generated from hundreds of actors under varied recording conditions, providing a challenging testbed for robustness analysis (Dolhansky et al. 2020). Other datasets, such as DeeperForensics-1.0 (Jiang et al. 2020), WildDeepfake (Zi et al. 2020) and DF40 (Yan et al. 2024), further increase forgery diversity or incorporate unconstrained real-world content, but often lack pixel-level manipulation masks, which limits their direct applicability to fine-grained spatial explainability evaluation.

Considering the objectives of this study, two datasets are selected for experimental evaluation. The DeepFakes subset of FaceForensics++ (FF-DF) is adopted as the primary dataset for both training and evaluation, as it offers paired real and fake samples with pixel-level manipulation masks and consistent preprocessing, enabling quantitative assessment of spatial alignment between explanation maps and manipulated regions (Rössler et al. 2019). To complement this controlled in-domain setting, Celeb-DF-v2 is used exclusively as a cross-dataset benchmark to assess the generalisation of explainability methods under domain shift. Celeb-DF-v2 provides matched real and manipulated samples while exhibiting higher synthesis quality and increased visual realism, thereby allowing explanations generated by models trained on FF++ to be evaluated under more challenging and less constrained real-world conditions (Li, Yang, et al. 2020). A summary of representative deepfake datasets and their key characteristics is provided in Table 2.

3.3 | Evaluation Pipeline for Explainable Deepfake Detection

To systematically evaluate the quality of post hoc visual explanations in the context of deepfake detection, this study designs a unified explainability evaluation pipeline, as illustrated in Figure 1. The pipeline follows a sequential and modular structure. First, a deepfake detection model is applied to input images or video frames to produce classification outputs, which constitute the decision basis to be explained. Second, post hoc explainability methods are employed to generate attribution or saliency maps that highlight the spatial evidence contributing to the detector's prediction. Third, the generated explanations are systematically evaluated through four complementary evaluation modules, each targeting distinct functional aspects of explanation quality.

Specifically, the evaluation stage comprises four complementary modules. Mask Alignment is introduced first to assess the spatial correspondence between explanation maps and ground-truth manipulated regions, and is described in detail in Section 3.3.1. Incremental Deletion is then employed to evaluate predictive faithfulness by measuring changes in model confidence under progressive removal of attributed evidence, as detailed in Section 3.3.2. Covariate Homogeneity follows to examine the consistency of explanations across semantically or temporally related samples, with the corresponding methodology presented in Section 3.3.3. Finally, Temporal Continuity is used to measure the stability of explanations across adjacent frames in a video, and its formulation is described in Section 3.3.4.

Each evaluation module is explicitly aligned with a specific subset of the Co-12 property framework, and together they cover multiple dimensions of explanation quality. This design enables different functional properties of explainability to be isolated and quantified in a principled and reproducible manner. In this study, seven Co-12 properties are selected to represent the relevant dimensions of explanation quality. Their definitions are summarised as follows (Nauta et al. 2023):

TABLE 2 | Overview of representative public deepfake datasets.

Dataset	Modality	Fake samples	Identities	Annotations	Link
FaceForensics (FF) (Rössler et al. 2018)	Video	1000	977	Video-level	https://github.com/ondyari/FaceForensics
FaceForensics++ (FF++) (Rössler et al. 2019)	Video	1000	977	Pixel-level masks	https://github.com/ondyari/FaceForensics
DeepFake Detection (DFD) (Ai 2019)	Video	3000	28	Video-level	https://github.com/ondyari/FaceForensics/tree/master/dataset
Celeb-DF (Li, Yang, et al. 2020)	Video	795 + 590	13 + 59	Video-level	https://github.com/yuezunli/celeb-deepfakeforensics
Celeb-DF-v2 (Li, Gao, et al. 2020)	Video	5639	59	Video-level	https://github.com/yuezunli/celeb-deepfakeforensics
DeepFake Detection Challenge (DFDC) (Dolhansky et al. 2020)	Video	5214	66	Video-level	https://www.kaggle.com/c/deepfake-detection-challenge
UADFV (Yang et al. 2019)	Video	49	—	Video-level	https://docs.google.com/forms/d/e/1FAIpQLScKPoOv15TIZ9Mn0nGScIVgKRM9tFWOmjh9eHKx57Yp-XcnxA/viewform
Deepfake-TIMIT (Korshunov and Marcel 2018)	Video	620	64	Video-level	https://www.idiap.ch/dataset/deepfaketimit
DeeperForensics-1.0 (Jiang et al. 2020)	Video	60,000	100	Video-level	https://github.com/EndlessSora/DeeperForensics-1.0
Fake Faces in the Wild (FFIW) (Khodabakhsh et al. 2018)	Image	150	—	Image-level	https://github.com/tfzhou/FFIW

- **Correctness**: The degree to which an explanation faithfully reflects the actual decision process of the underlying model.
- **Completeness**: The extent to which an explanation accounts for the full range of factors influencing the model's prediction, rather than only a subset
- **Compactness**: The parsimony of an explanation, describing how concisely the relevant information is presented without unnecessary detail.
- **Covariate Completeness**: The consistency of an explanation with respect to covariate structure, reflecting whether explanations remain coherent across samples with similar properties.
- **Continuity**: The smoothness and stability of explanations, such that similar inputs yield similar explanatory outputs
- **Composition**: The organisational quality of an explanation, referring to how well spatial and temporal elements are integrated into a coherent presentation.
- **Coherence**: The plausibility of an explanation in light of prior knowledge or ground-truth evidence, indicating whether it aligns with what is expected or reasonable.

The mapping between each evaluation method and the corresponding Co-12 properties is presented in Table 3.

3.3.1 | Mask Alignment

This evaluation component measures the spatial correspondence between model-generated explanation maps and ground-truth manipulated regions in deepfake videos. In imaging-based explainability, such correspondence is commonly quantified using overlap-based metrics such as Intersection over Union (IoU), outside–inside relevance ratio, localisation error, or pointing game accuracy (Huang and Li 2020; Nam et al. 2020).

The objective is to quantify how well highlighted regions overlap with annotated tampered areas defined by pixel-level masks. Since explanation methods produce saliency maps with different sparsity, scale and distribution, the evaluation procedure includes steps to handle these variations. Potential annotation inaccuracies near boundaries and resolution mismatches are addressed through resolution harmonisation, boundary adjustment and scale normalisation.

Attribution maps of fake images are first normalised to the $[0, 1]$ range. Reference masks are binarised and dilated by one iteration with a 3×3 kernel to reduce sensitivity to boundary discrepancies. When the resolutions of maps and masks differ, bilinear resizing is applied. To mitigate scale effects, the attribution mass is adjusted to match the area of the reference mask.

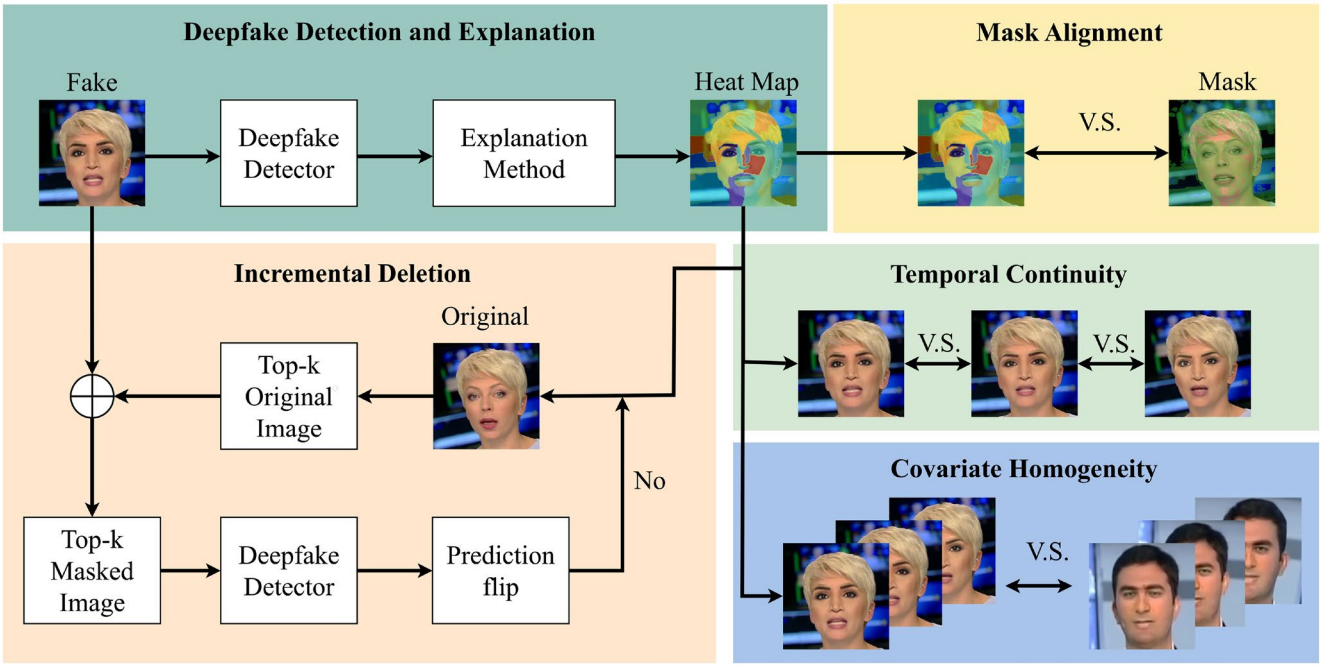

FIGURE 1 | Overview of the explainability evaluation pipeline for deepfake detection. The pipeline consists of three stages: Deepfake detection, post hoc explanation generation and modular evaluation of explanation quality.

Two overlap-based metrics are used: *Soft-IoU* and *Top-k IoU*. The Soft-IoU measures the continuous overlap between the normalised attribution map $\hat{E}$ and the binary mask M:

$$\mathrm{IoU} = \frac{\left\langle \hat{E}, M \right\rangle}{\| \hat{E} \|_1 + \| M \|_1 - \left\langle \hat{E}, M \right\rangle},$$

where $\hat{E}$ is a normalised, non-negative explanation and $M \in \{0,1\}^{H \times W}$ is the dilated ground-truth mask.

To evaluate alignment under different support levels, *Top-k IoU* is defined. For a given support percentage k, the top $K = \lfloor k / 100 \cdot N \rfloor$ values in $\hat{E}$ are retained to form a sparsified map $\hat{E}^{\mathrm{soft}}_{(k)}$:

$$\hat{E}^{\mathrm{soft}}_{(k)}(i) = \begin{cases} \hat{E}(i), & i \in S_k, \\ 0, & \text{otherwise}, \end{cases}$$

where S_k denotes the set of indices corresponding to the top K values. The Top-k IoU is then computed as:

$$\mathrm{IoU}^{\mathrm{soft}}_k = \frac{\left\langle \hat{E}^{\mathrm{soft}}_{(k)}, M \right\rangle}{\| \hat{E}^{\mathrm{soft}}_{(k)} \|_1 + \| M \|_1 - \left\langle \hat{E}^{\mathrm{soft}}_{(k)}, M \right\rangle}.$$

Alignment scores are computed per sample and aggregated across methods. Soft-IoU and Top-k IoU reduce dependence on binary thresholding and enable comparison across attribution maps with different distributions.

Precision and Recall are not included, as they are highly sensitive to attribution scale and density and may reflect methodological preferences rather than true spatial alignment. IoU-based measures instead provide a more stable basis for comparison.

3.3.2 | Incremental Deletion

This component evaluates the predictive faithfulness of attribution maps by incrementally masking regions identified as important or unimportant and measuring the resulting effect on the model's confidence (Nauta et al. 2023). If masking highly attributed regions leads to a substantial prediction drop, the attribution is considered faithful, whereas masking low-attribution regions should have limited impact.

Each attribution map is normalised to the [0, 1] range. A set of masking ratios $K = \{0.5,1,1.5,2, \ldots, 90\}$ is defined. For each $k \in K$, the top-k % and bottom-k % pixels define two complementary binary masks: PGI (globally important) and PGU (globally unimportant).

Masked regions are replaced using pristine content from the corresponding unmanipulated image, ensuring that prediction changes reflect evidence removal rather than masking artefacts (Tsigos et al. 2024). Alternative masking strategies (mean and blur) are also implemented.

For each masking level, the model's predicted fake-class probability is recorded, forming deletion curves for PGI and PGU. The area under the curve (AUC) is computed using the trapezoidal rule, and a normalised AUC is obtained by dividing each trajectory by its starting value.

Several metrics are derived:

TABLE 3 | Co-12 attributes covered by each evaluation module.

Method	Correctness	Completeness	Compactness	Covariate Compl.	Continuity	Composition	Coherence
Mask alignment						✓	✓
Incremental deletion	✓	✓	✓				
Covariate homogeneity				✓			✓
Temporal continuity					✓	✓	✓

- **Correctness**: $\Delta\text{AUC}_{\text{PGI}}$, the AUC reduction when important regions are masked.
- **Completeness**: terminal probability p_L^{PGI} and total drop $\Delta p = p_0 - p_L^{\text{PGI}}$.
- **Compactness**: k_{flip}, the smallest masking ratio at which confidence crosses a threshold.

This design contrasts important and unimportant regions and normalises trajectories to support consistent comparison across samples.

3.3.3 | Covariate Homogeneity

This component evaluates the consistency of attribution maps across semantically or temporally related samples, particularly frames from the same video. Prior work has assessed such consistency using measures such as cluster purity or variance across groups (Shi et al. 2020; Lakkaraju and Leskovec 2016).

Each attribution map is normalised, resized to $r \times r$, and flattened into a vector. Vectors are grouped by video identity, and groups with fewer than three samples are excluded.

Within-group and between-group variances are computed, and their ratio defines:

$$F = \frac{\text{inter}}{\text{intra} + \epsilon},$$

with an adjusted score

$$F_{\text{adj}} = F \cdot (\overline{n} - 1),$$

where $\overline{n}$ is the average group size.

Cosine similarity between each attribution vector and its group mean is also computed.

Reported statistics include group counts, variances, F-scores, adjusted F_{adj}, effective group size and mean cosine similarity, providing a concise description of intra-video attribution consistency.

3.3.4 | Temporal Continuity

This component evaluates whether attribution maps vary smoothly over time by measuring their similarity across adjacent frames within the same video. Since visual content typically changes gradually, stable explanations are expected to remain consistent across neighbouring frames. Prior work measures explanation stability using cosine similarity or related metrics (Plumb et al. 2020; Chu et al. 2019).

Frames are temporally ordered, and adjacent pairs within a maximum gap Δt are constructed. Attribution maps are normalised, resized to $r \times r$, flattened and cosine similarity is computed:

$$\cos(v_a, v_b) = \frac{\langle v_a, v_b \rangle}{\| v_a \| \, \| v_b \|}.$$

The mean cosine similarity over all valid pairs is reported, together with the number of evaluated pairs and hyperparameters r and Δt.

4 | Comparative Evaluation

4.1 | Qualitative Results

Figures 2 and 3 present a qualitative comparison of explanation methods on the FF-DF dataset. Figure 2 corresponds to manipulated inputs correctly identified as fake by the detector, while Figure 3 corresponds to pristine inputs correctly identified as authentic.

For manipulated inputs, several methods are able to highlight regions that correspond to the manipulated areas. Vanilla Gradient, Integrated Gradients and SmoothGrad emphasise facial regions around the mouth, chin and cheeks. Guided Backprop and Deconvolution generate fine-grained but noisy patterns. Grad-CAM, Grad-CAM++, Score-CAM and CAM produce concentrated activations in central facial regions, which capture major manipulations but do not consistently cover peripheral artefacts. Occlusion Sensitivity, KernelSHAP and LIME provide block-structured highlights that frequently overlap with manipulated regions. Layer-wise Relevance

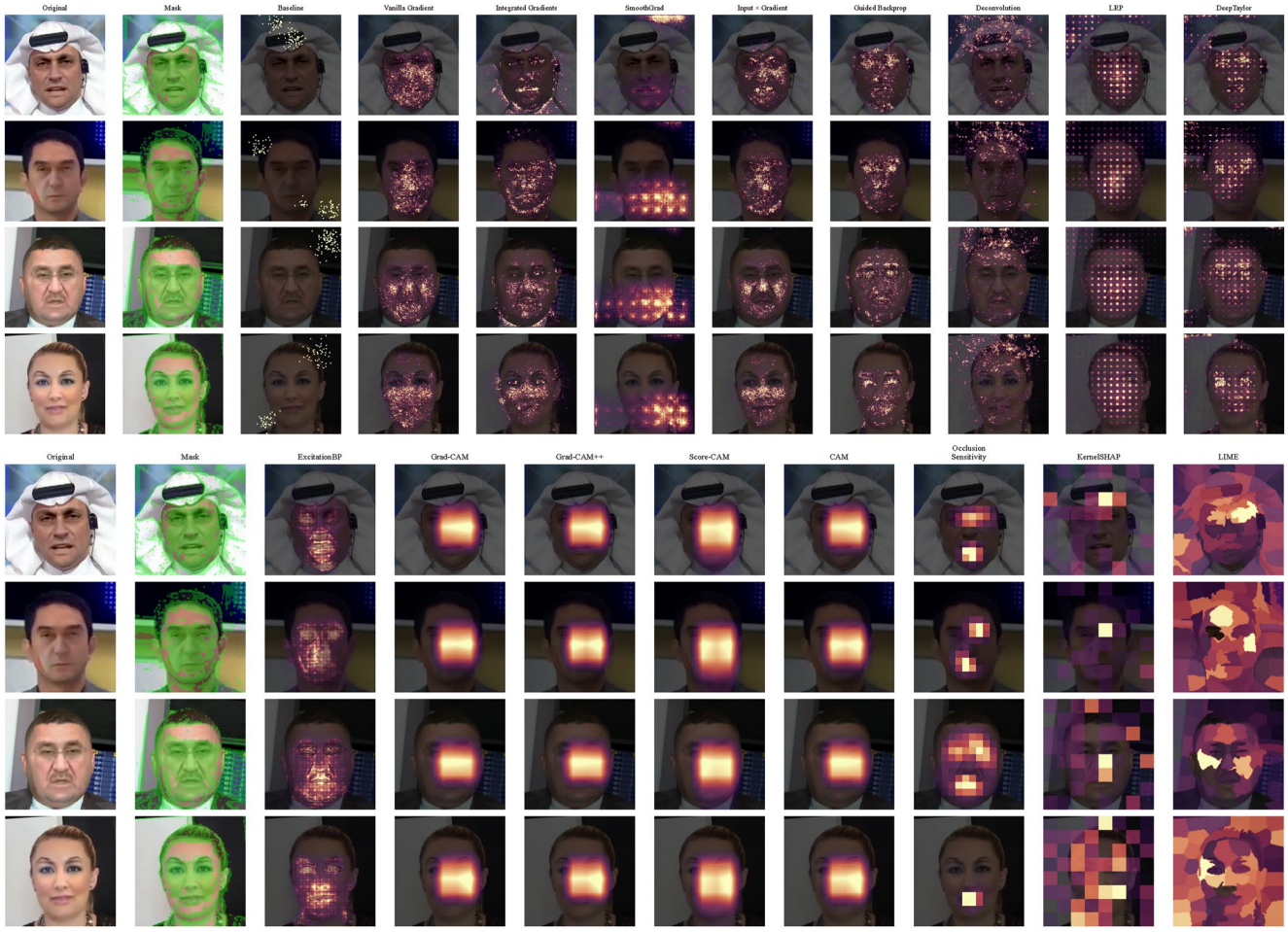

FIGURE 2 | Qualitative comparison of explanation methods on manipulated samples.

Propagation and Deep Taylor Decomposition yield fragmented saliency maps that are less consistent with the ground-truth manipulated areas.

For pristine inputs, explanation methods exhibit different behaviour. Vanilla Gradient, Integrated Gradients and Guided Backprop distribute saliency broadly across facial features including the eyes and mouth, despite the absence of manipulation. Grad-CAM, Grad-CAM++, Score-CAM and CAM concentrate activations in central facial regions, which reflects the discriminative focus of the detector rather than evidence of manipulation. Occlusion Sensitivity, KernelSHAP and LIME produce sparse and localised responses, which align more closely with the expectation that no manipulation-specific regions should be present. Layer-wise Relevance Propagation and Deep Taylor Decomposition again generate fragmented and noisy saliency maps, which complicates interpretation.

Overall, gradient-based methods provide fine-grained but noisy visualisations, CAM-based methods generate concentrated yet sometimes incomplete activations, and perturbation-based methods yield more interpretable localised responses. These findings demonstrate the importance of qualitative inspection for understanding the interpretability of deepfake detection models.

4.2 | Quantitative Results

4.2.1 | Mask Alignment

This component evaluates the spatial alignment between explanation maps and ground-truth manipulated regions. Figure 4 summarises the distribution of IoU scores across explanation methods on FF-DF and Celeb-DF-v2, together with their dataset-level averages. A clear performance gap is observed between the two datasets: for nearly all methods, IoU scores on FF-DF are consistently higher than those on Celeb-DF-v2. This discrepancy is expected, as the detection model is trained on FaceForensics++ Deepfakes, making FF-DF an in-domain setting, whereas Celeb-DF-v2 represents a more challenging cross-domain scenario with higher visual realism and different manipulation artefacts.

Within the CAM category, clear performance stratification can be observed. Grad-CAM++ and CAM achieve the strongest spatial alignment, with dataset-level average IoU values exceeding 0.84 and consistently high scores on both FF-DF and Celeb-DF-v2. Other CAM-based variants exhibit more moderate alignment. In particular, Score-CAM shows competitive performance on FF-DF but experiences a larger reduction on Celeb-DF-v2, resulting in a lower dataset-average IoU and placing it

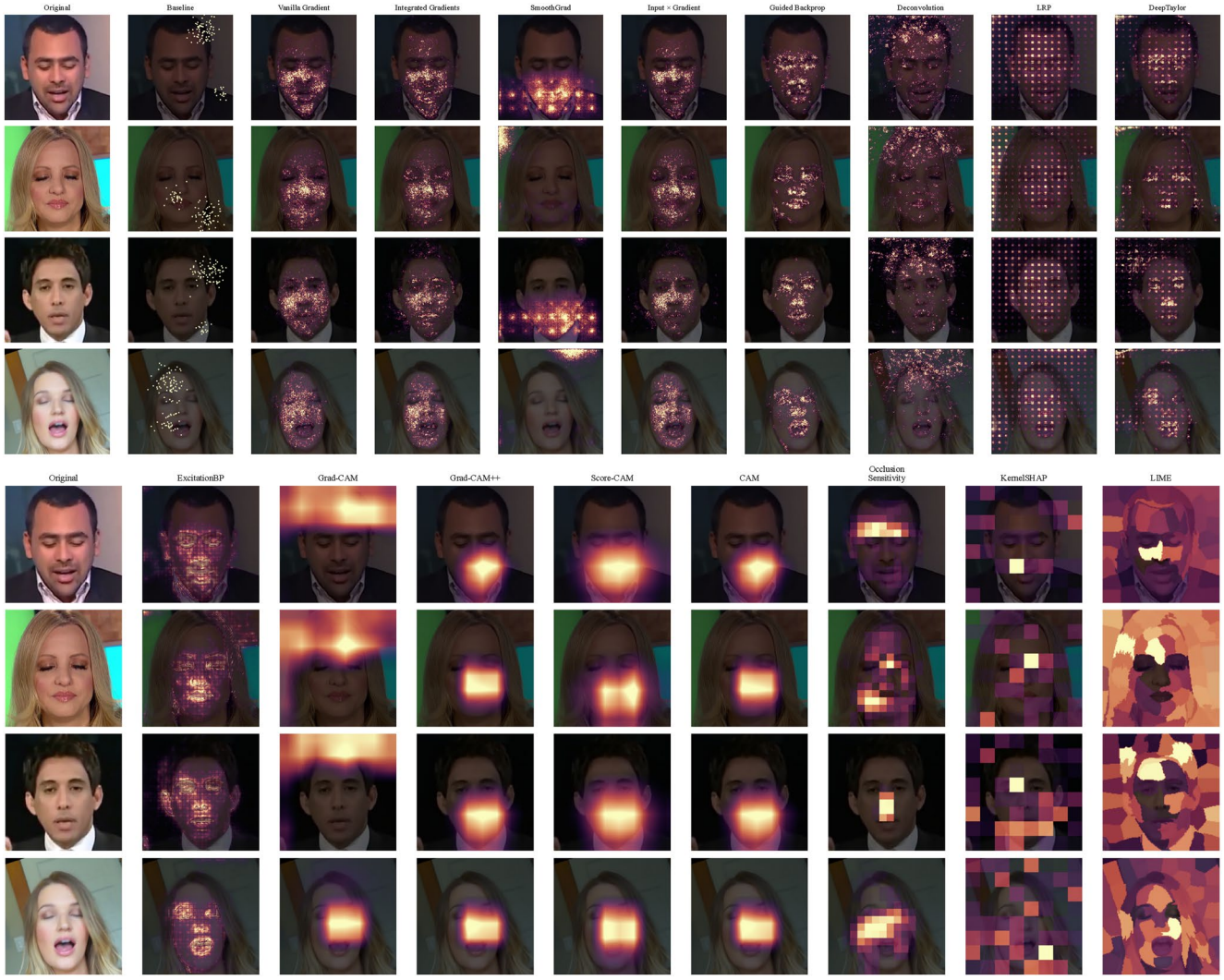

FIGURE 3 | Qualitative comparison of explanation methods on pristine samples.

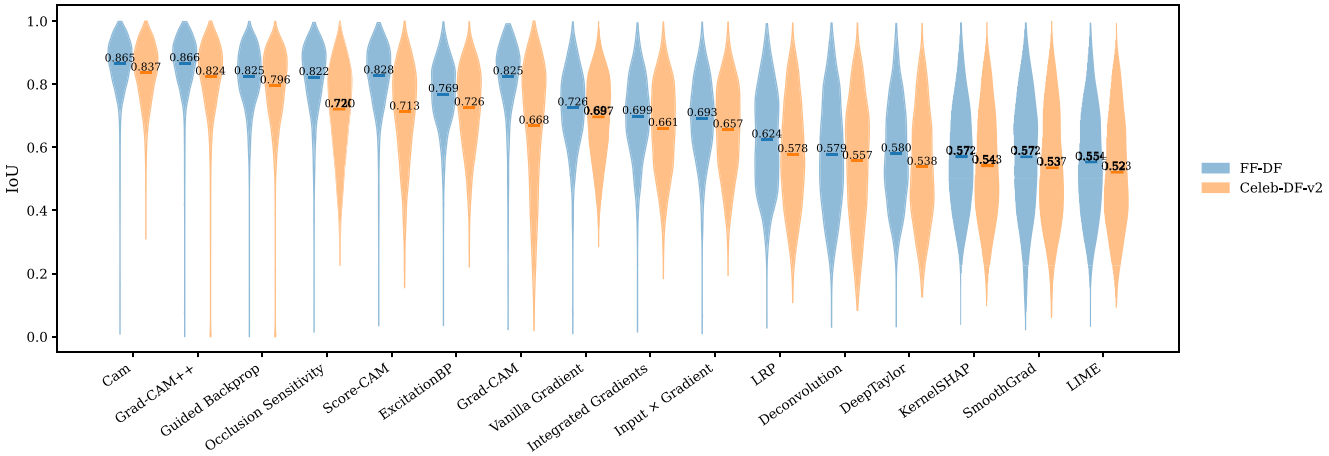

FIGURE 4 | Distribution of IoU scores across methods.

below the leading CAM variants in terms of robustness under cross-domain evaluation. Overall, CAM-derived methods remain among the most effective approaches for mask alignment, but their stability varies depending on how class-specific activations are constructed and aggregated.

Gradient-based attribution methods show intermediate spatial alignment, with noticeable variation across specific techniques. Guided Backprop achieves the highest alignment within this family, reaching a dataset-level average IoU of ~0.81 (0.825 on FF-DF and 0.796 on Celeb-DF-v2), indicating relatively stable

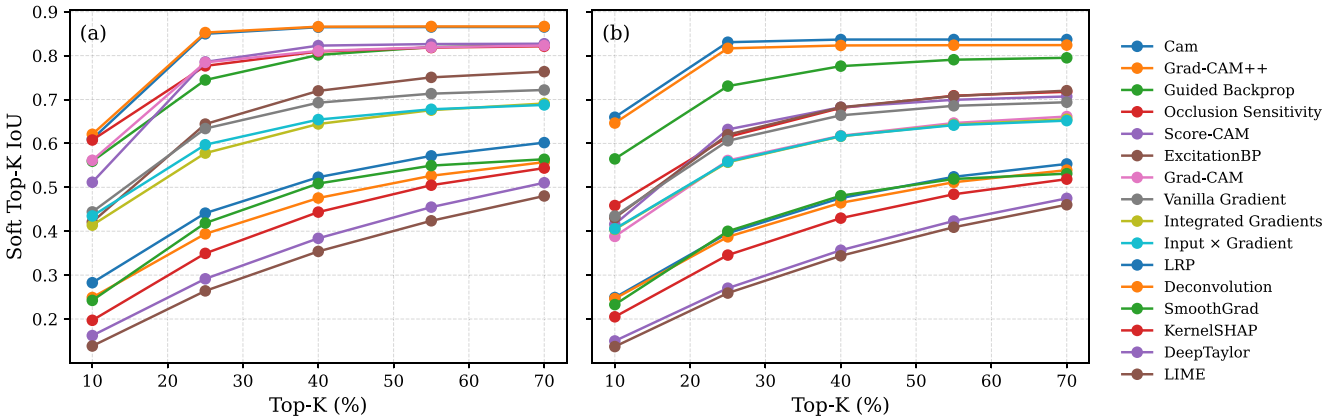

FIGURE 5 | Top-*k* IoU curves for explanation methods. Subfigure (a) corresponds to FF-DF, and subfigure (b) corresponds to Celeb-DF-v2.

correspondence with manipulated regions across domains. In contrast, other gradient-based approaches exhibit lower and more dispersed IoU values. Vanilla Gradient attains an average IoU of 0.712, while Integrated Gradients and Input×Gradient further decrease to around 0.68 and 0.67, respectively. Across datasets, all gradient-based methods demonstrate a consistent reduction in IoU on Celeb-DF-v2 compared to FF-DF, reflecting increased difficulty in localising manipulations under cross-domain conditions. This trend is also evident in the Top-*k* IoU curves, where Guided Backprop reaches higher saturation levels, whereas the remaining gradient-based methods plateau earlier at moderate overlap levels, indicating limited recovery of spatial alignment when progressively including less salient pixels.

Perturbation-based methods display a wider spread of spatial alignment performance, largely determined by the granularity of the perturbation strategy. Occlusion Sensitivity achieves the strongest alignment within this group, with a dataset-level average IoU of approximately 0.77 (0.822 on FF-DF and 0.720 on Celeb-DF-v2), placing it clearly above other perturbation-based approaches. Its relatively smaller performance drop across datasets suggests that region-wise masking remains effective for identifying manipulated areas even under domain shift. In contrast, KernelSHAP and LIME consistently obtain lower IoU scores, with dataset-level averages of 0.56 and 0.54, respectively. Their performance declines further on Celeb-DF-v2, where IoU values fall close to or below 0.52. The corresponding Top-*k* IoU curves rise more slowly and converge at lower levels, indicating that increasing the number of selected regions does not substantially improve overlap with ground-truth masks, consistent with the coarse superpixel-based perturbations employed by these methods.

Figure 5 further illustrates these trends through Top-*k* IoU curves. CAM and Grad-CAM++ exhibit rapid growth and early saturation above 0.8, indicating that their most salient regions strongly overlap with manipulated areas even at small *k*. Guided Backprop and Occlusion Sensitivity show stable but noticeably lower saturation levels, while Score-CAM grows more gradually and plateaus below the CAM variants, particularly in the cross-domain Celeb-DF-v2 setting.

Pure gradient-based methods plateau earlier and at substantially lower IoU values, suggesting that increasing the number of selected pixels does not substantially recover alignment with manipulation masks. KernelSHAP and LIME demonstrate the slowest growth and lowest convergence levels, reflecting their limited ability to localise fine-grained forgery regions under Top-*k* selection.

The observed differences across method families are closely linked to their attribution mechanisms. CAM-based methods aggregate class-specific feature activations, resulting in compact and semantically coherent saliency maps that align well with manipulated facial regions. Gradient-based methods are more sensitive to local intensity changes and edges, which may partially coincide with manipulation boundaries but often extend into irrelevant areas.

Perturbation-based methods exhibit contrasting behaviour. Occlusion Sensitivity benefits from directly measuring prediction changes under region removal, leading to comparatively stronger alignment than other perturbation approaches. In contrast, KernelSHAP and LIME rely on superpixel-based abstractions, which produce spatially coarse and irregular attribution regions that fail to capture fine manipulation boundaries, explaining both their low IoU scores and slow Top-*k* growth.

Overall, Mask Alignment results indicate that explanation faithfulness is strongly influenced by both the attribution mechanism and the domain consistency between training and evaluation data. While CAM-based methods remain robust under moderate domain shift, most other methods experience a noticeable degradation on Celeb-DF-v2, highlighting the importance of considering dataset bias when interpreting spatial explanation quality.

4.2.2 | Incremental Deletion

This component evaluates the faithfulness of explanation methods through incremental deletion, where pixels are progressively removed according to their attributed importance. The analysis is conducted under two complementary protocols. In the PGI setting, pixels ranked as important are removed sequentially until the selected set is exhausted, revealing how strongly explanations capture decision-critical evidence. In the PGU setting, pixels ranked as unimportant are removed,

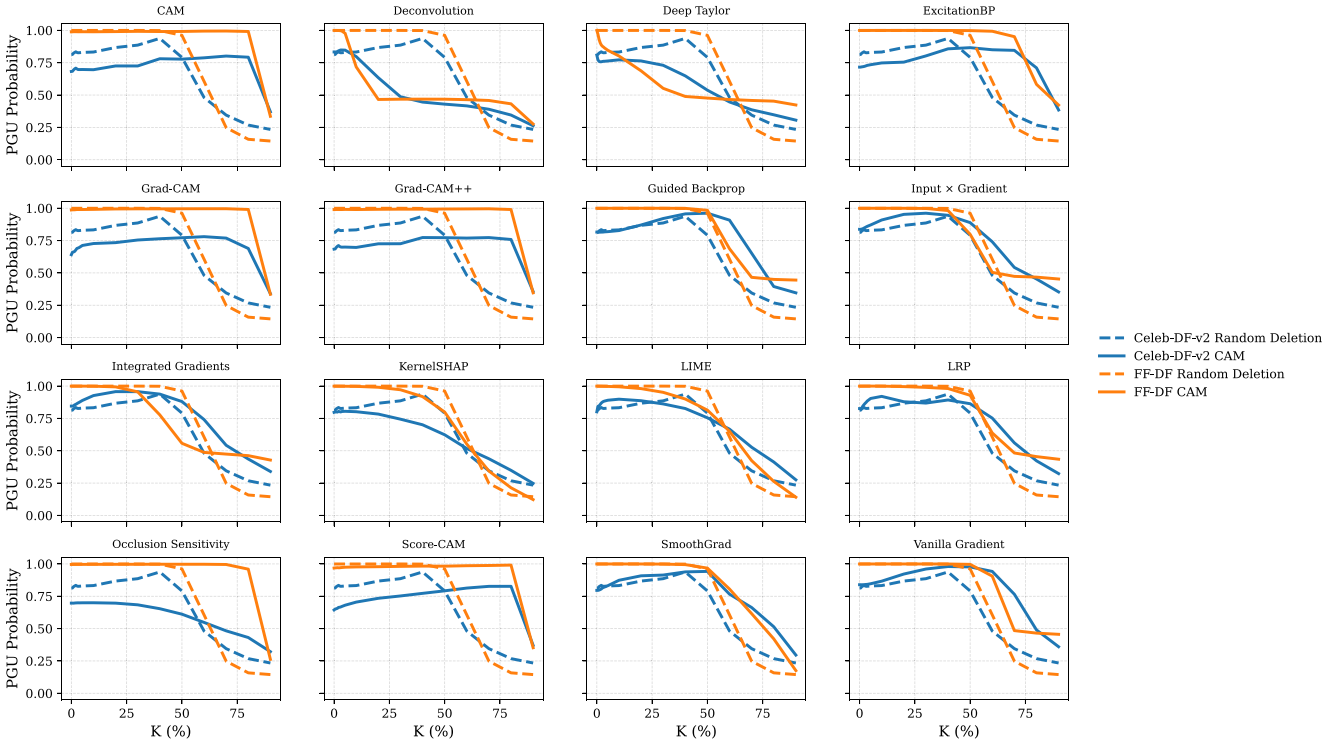

FIGURE 6 | Probability trajectories under globally unimportant (PGU) deletion across methods, compared with random deletion.

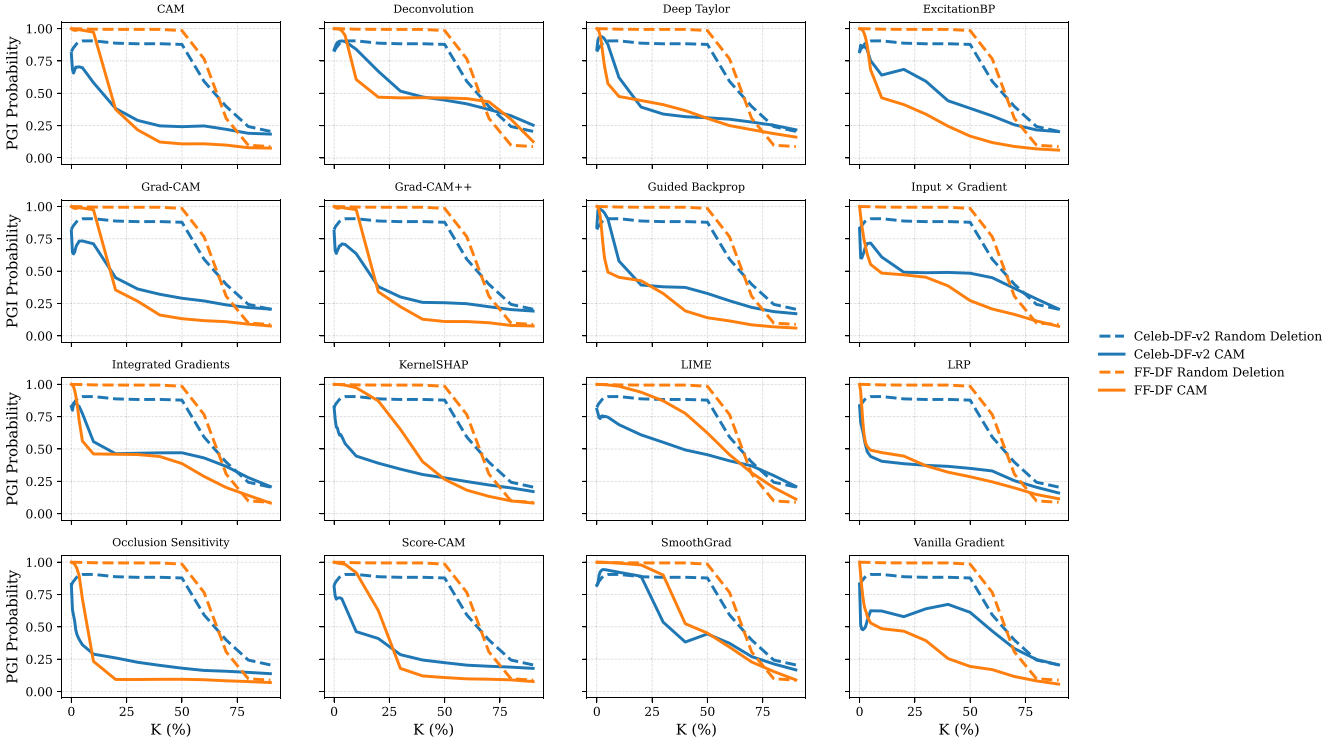

FIGURE 7 | Probability trajectories under globally important (PGI) deletion across methods, compared with random deletion.

providing a qualitative reference for how model confidence behaves when irrelevant regions are perturbed. As a reference, we also report random deletion, where the same proportion of pixels is removed uniformly at random, providing an importance-agnostic control. Faithfulness is quantified using three attributes derived from PGI deletion: correctness, output-completeness and compactness, which together characterise the causal relevance, sufficiency and spatial concentration of explanations.

Figures 6 and 7 illustrate representative confidence trajectories on FF-DF and Celeb-DF-v2 under PGI and PGU deletion. Under PGI deletion, most methods exhibit a monotonic decline in confidence as increasingly important pixels are removed, but

TABLE 4 | Evaluation of correctness, output-completeness and compactness.

Method	FF-DF				Celeb-DF-v2			
(lr)2–5 (lr)6–9	ΔAUC_{PGI}	Δ P	Flip rate	k_{flip}	ΔAUC_{PGI}	Δ P	Flip rate	k_{flip}
Random deletion	0.00	0.67	0.83	66.23	0.00	0.32	0.67	59.99
Occlusion sensitivity	0.38	0.68	0.93	10.35	0.28	0.34	0.81	6.05
Deep taylor	0.36	0.65	0.96	13.14	0.14	0.22	0.66	18.91
Guided backprop	0.35	0.69	0.98	9.77	0.13	0.34	0.71	22.18
Grad-CAM++	0.27	0.68	0.78	26.33	0.19	0.33	0.67	12.40
CAM	0.27	0.68	0.78	26.12	0.19	0.33	0.67	12.36
Score-CAM	0.28	0.69	0.78	25.42	0.18	0.32	0.66	12.38
KernelSHAP	0.22	0.70	0.89	38.64	0.24	0.33	0.77	15.21
LRP	0.29	0.48	0.96	9.66	0.12	0.21	0.73	11.57
ExcitationBP	0.26	0.67	0.89	17.78	0.12	0.33	0.64	22.87
Grad-CAM	0.23	0.67	0.76	27.77	0.14	0.32	0.65	15.49
Deconvolution	0.26	0.70	0.97	21.15	0.08	0.24	0.66	34.51
Input × Gradient	0.26	0.59	0.96	12.92	0.03	0.28	0.69	17.34
Vanilla gradient	0.26	0.65	0.96	12.14	0.04	0.32	0.72	15.67
Integrated gradients	0.26	0.56	0.96	11.91	0.03	0.24	0.64	19.67
SmoothGrad	0.07	0.62	0.74	50.91	0.04	0.29	0.62	31.46
LIME	0.00	0.48	0.64	50.06	−0.01	0.18	0.56	25.15

Note: The data highlighted in red represent the optimal value for each corresponding column; all Δ-based metrics are reported with two decimal places.

the rate and onset of this decline vary substantially across methods and datasets. Under PGU deletion, confidence generally remains stable at early deletion stages and decays only when a large fraction of pixels is removed, serving as a sanity check that explanations do not misidentify irrelevant regions as important. These qualitative patterns are summarised quantitatively by the metrics reported in Table 4.

Correctness is reflected by the extent to which removing important pixels suppresses the model's confidence throughout the deletion process, as measured by ΔAUC_{PGI}. Occlusion Sensitivity achieves the highest correctness on both datasets, with ΔAUC_{PGI} values of 0.38 on FF-DF and 0.28 on Celeb-DF-v2, indicating that the regions it identifies are the most causally influential in aggregate. Deep Taylor and Guided Backprop also exhibit strong correctness on FF-DF, reaching ΔAUC_{PGI} values around 0.35, but their correctness degrades markedly on Celeb-DF-v2, where the corresponding values drop to ~0.14 and 0.13. CAM-based methods show more moderate correctness, with ΔAUC_{PGI} values clustered around 0.27 on FF-DF and around 0.18–0.19 on Celeb-DF-v2, reflecting consistent yet less pronounced confidence suppression under PGI deletion.

Output-completeness is captured by the overall probability drop ΔP under PGI deletion, which reflects whether the set of important pixels identified by an explanation is sufficient to account for the model's final decision. On FF-DF, KernelSHAP achieves the largest ΔP value of 0.70, followed closely by Deconvolution and Guided Backprop, indicating that their selected pixels collectively remove a large portion of the model's confidence by the end of the deletion process. On Celeb-DF-v2, the highest ΔP values are observed for Occlusion Sensitivity and Guided Backprop, both around 0.34. The divergence between methods that maximise ΔAUC_{PGI} and those that maximise ΔP suggests that correctness and output-completeness capture complementary aspects of faithfulness: some methods induce sustained confidence reduction throughout deletion, while others achieve a large terminal drop without dominating the entire trajectory.

Compactness describes how concentrated the explanation is and is jointly characterised by the flip rate and the flip point k_{flip} under PGI deletion. The flip rate measures how often the prediction is overturned when important pixels are progressively removed, while k_{flip} indicates the proportion of pixels required to trigger this reversal. A higher flip rate combined with a smaller k_{flip} therefore indicates a more compact explanation. On FF-DF, Guided Backprop and LRP exhibit the strongest compactness, achieving high flip rates together with k_{flip} values below 10%, meaning that removing a very small fraction of top-ranked pixels is sufficient to change the prediction. Occlusion Sensitivity also demonstrates strong compactness, with a flip point around 10%. In contrast, CAM-based methods require substantially larger deletion ratios, with k_{flip} values around 25%–27%, indicating more spatially distributed attributions. KernelSHAP and LIME show the weakest compactness, as reflected by large k_{flip} values exceeding 35% or even 50%.

Overall, FF-DF consistently yields stronger faithfulness characteristics than Celeb-DF-v2 across all three attributes. Higher ΔAUC_{PGI} and ΔP, together with smaller k_{flip} values, indicate that explanations align more closely with the detector's reasoning in the training domain. Celeb-DF-v2 introduces systematic attenuation across metrics, reflecting the impact of domain shift and increased manipulation realism. Importantly, the results reveal clear trade-offs between faithfulness attributes: Occlusion Sensitivity excels in correctness and cross-domain robustness, Guided Backprop and LRP produce highly compact explanations in-domain and CAM-based methods provide stable but less concentrated attribution. These trade-offs suggest that the suitability of an explanation method depends on which aspect of faithfulness is prioritised in forensic analysis.

4.2.3 | Covariate Homogeneity

This component evaluates the homogeneity of attribution maps by measuring their consistency within videos and separability across different videos. Table 5 summarises the homogeneity statistics across methods, where the data highlighted in red represent the optimal value for each corresponding column. Higher intra-group scores indicate stronger within-video consistency, while lower inter-group scores indicate better separation across videos. The adjusted score F_{adj} integrates both aspects to provide an overall measure of homogeneity.

Across all methods, a clear dataset-dependent trend is observed. For most approaches, homogeneity scores on FF-DF are consistently higher than those on Celeb-DF-v2, reflected by larger F_{adj} values and, in many cases, higher intra-group similarity. This difference can be attributed to the fact that the detector is trained on FaceForensics++ Deepfakes, making FF-DF an in-domain setting where attribution patterns are more stable across frames. In contrast, Celeb-DF-v2 introduces stronger domain shift and higher visual realism, which weakens both within-video consistency and cross-video separability.

Among all methods, ExcitationBP achieves the highest overall homogeneity on both datasets, with F_{adj} values of 36.36 on FF-DF and 26.91 on Celeb-DF-v2, as well as the largest average score across datasets. Its relatively low inter-group scores combined with stable intra-group similarity indicate that the resulting attribution maps are both consistent within videos and well separated across different videos. CAM also demonstrates strong homogeneity, reaching F_{adj} values above 36 on FF-DF and above 25 on Celeb-DF-v2, with balanced intra and inter statistics that remain relatively stable under cross-domain evaluation.

Several CAM-derived variants exhibit similar but slightly weaker behaviour. Grad-CAM++ and Grad-CAM achieve moderately high F_{adj} values on FF-DF (above 27) but experience a noticeable reduction on Celeb-DF-v2, where the scores drop to around 19–20. This reduction is accompanied by a marked increase in intra-group values on Celeb-DF-v2, suggesting that while frame-level similarity remains high, the separation between different videos becomes less pronounced under domain shift. Score-CAM shows a comparable

TABLE 5 | Homogeneity statistics across methods.

	FF-DF			Celeb-DF-v2			AVG		
Method	F_{adj}	**Intra** ($\times 10^{-2}$)	**Inter** ($\times 10^{-2}$)	F_{adj}	**Intra** ($\times 10^{-2}$)	**Inter** ($\times 10^{-2}$)	F_{adj}	**Intra** ($\times 10^{-2}$)	**Inter** ($\times 10^{-2}$)
ExcitationBP	36.36	0.096	0.110	26.91	0.120	0.106	31.64	0.108	0.108
CAM	36.04	0.300	0.291	25.34	0.428	0.349	30.69	0.364	0.320
SmoothGrad	33.15	0.206	0.220	22.30	0.192	0.140	27.73	0.199	0.180
Grad-CAM++	32.70	0.369	0.307	19.12	0.613	0.374	25.91	0.491	0.340
Grad-CAM	27.20	1.214	1.171	20.47	2.869	1.913	23.83	2.042	1.542
Occlusion Sensitivity	26.95	0.888	0.649	15.38	1.208	0.600	21.17	1.048	0.624
LRP	28.32	0.149	0.092	10.12	0.265	0.086	19.22	0.207	0.089
Score-CAM	21.87	0.349	0.236	13.95	1.035	0.471	17.91	0.692	0.353
LIME	16.73	3.012	1.625	11.73	3.387	1.295	14.23	3.200	1.460
DeepTaylor	13.89	0.162	0.071	10.98	0.262	0.094	12.44	0.212	0.082
Guided Backprop	9.33	0.053	0.016	6.34	0.070	0.014	7.84	0.062	0.015
Integrated Gradients	9.09	0.092	0.027	6.36	0.104	0.022	7.72	0.098	0.024
Input × Gradient	8.17	0.113	0.030	6.84	0.112	0.025	7.50	0.112	0.027
Vanilla Gradient	6.06	0.145	0.028	5.47	0.139	0.025	5.76	0.142	0.027
Deconvolution	3.28	0.121	0.013	2.80	0.122	0.011	3.04	0.121	0.012
KernelShap	1.96	4.834	0.304	2.08	4.474	0.301	2.02	4.654	0.303

Note: The data highlighted in red represent the optimal value for each corresponding column. Intra- and inter-group scores are reported in scientific notation as $\times 10^{-2}$.

pattern, with F_{adj} decreasing from 21.87 on FF-DF to 13.95 on Celeb-DF-v2.

Gradient-based methods display more heterogeneous homogeneity characteristics. Methods such as Guided Backprop, Integrated Gradients and Input×Gradient achieve relatively low F_{adj} values on both datasets, remaining below 10 on FF-DF and dropping further on Celeb-DF-v2. Their intra-group scores are consistently small, indicating limited frame-level consistency, while inter-group scores are also small, reflecting weak discrimination across videos. SmoothGrad stands out within this group, achieving a substantially higher F_{adj} of 33.15 on FF-DF and 22.30 on Celeb-DF-v2, which indicates improved stability and separation compared to other gradient-based approaches.

Perturbation-based methods show the largest variability across datasets. Occlusion Sensitivity reaches relatively high homogeneity on FF-DF, with F_{adj} close to 27, but its score decreases to around 15 on Celeb-DF-v2, reflecting reduced consistency under domain shift. LIME exhibits high intra-group scores on both datasets, indicating strong within-video similarity, but its inter-group scores remain comparatively large, resulting in moderate F_{adj} values and limited separation across videos. KernelSHAP achieves the lowest F_{adj} among all methods on both datasets, despite having the highest intra-group scores, because its inter-group scores are also large, indicating that attribution patterns are similar across different videos and thus lack discriminative power.

Overall, the homogeneity analysis reveals that methods differ substantially in how they balance within-video consistency and cross-video separability. FF-DF consistently yields stronger homogeneity than Celeb-DF-v2, highlighting the impact of domain alignment on attribution stability. Methods such as ExcitationBP and CAM provide robust and well-balanced homogeneity across datasets, while gradient-based methods tend to produce less stable attribution patterns. Perturbation-based methods exhibit either strong intra-video similarity or weak inter-video separation, but rarely both simultaneously. These results demonstrate that homogeneity is highly sensitive to dataset characteristics and should be interpreted jointly with other evaluation criteria when assessing explanation quality.

4.2.4 | Temporal Continuity

This component evaluates the temporal stability of explanation methods across consecutive video frames. Results are summarised in Table 6, where the data highlighted in red represent the optimal value for each corresponding column. Temporal continuity is assessed from two complementary perspectives: cosine similarity, which measures overall frame-to-frame consistency of attribution maps, and IoU@10/20/30, which quantify spatial overlap of the most salient regions retained at different sparsity levels.

Across all methods, temporal continuity is consistently stronger on FF-DF than on Celeb-DF-v2. Most approaches exhibit higher cosine similarity and larger IoU@k values on FF-DF, indicating that attribution patterns are more stable across frames in the training-domain setting. In contrast, Celeb-DF-v2 introduces noticeable degradation across metrics, reflecting the increased temporal variability of explanations under domain shift and higher visual realism.

CAM-based methods achieve the strongest temporal continuity on both datasets. CAM and Score-CAM reach the highest cosine similarity values on FF-DF (both around 0.97) and maintain high consistency on Celeb-DF-v2 (above 0.92). These methods also dominate the IoU@k metrics, with CAM achieving the best IoU@10 and IoU@20 on both datasets and Score-CAM attaining the highest IoU@30 on FF-DF. The relatively small performance gap between FF-DF and Celeb-DF-v2 for CAM-based methods indicates that class-activation mechanisms preserve attribution structure more robustly across adjacent frames, even under domain shift.

Propagation-based and hybrid attribution methods demonstrate intermediate to high temporal stability. ExcitationBP shows balanced performance across datasets, with cosine similarity above 0.91 on both FF-DF and Celeb-DF-v2 and consistently high IoU@k values, indicating stable attribution propagation over time. Deep Taylor and LRP also achieve high cosine similarity, particularly on Celeb-DF-v2, but their IoU@k scores are notably lower than those of CAM-based methods, suggesting that although overall attribution patterns are temporally aligned, the spatial overlap of the most salient regions is less consistent across frames.

Gradient-based methods exhibit moderate temporal continuity. Vanilla Gradient, Integrated Gradients and Input × Gradient record cosine similarity values around 0.62–0.68 on both datasets, with corresponding IoU@k values remaining relatively low, especially at IoU@10. SmoothGrad outperforms other gradient-based variants, achieving cosine similarity close to 0.90 and substantially higher IoU@k scores, indicating that gradient smoothing effectively reduces frame-level noise and improves temporal stability.

Perturbation-based methods display the largest variability in temporal continuity. Occlusion Sensitivity yields moderate cosine similarity and IoU@k values, with a clear drop from FF-DF to Celeb-DF-v2, indicating reduced stability under domain shift. LIME exhibits relatively high cosine similarity on both datasets but much lower IoU@k values, suggesting that while attribution patterns remain globally similar across frames, the spatial locations of salient regions fluctuate. KernelSHAP consistently records the lowest cosine similarity and IoU@k scores across all datasets, indicating weak temporal continuity and unstable attribution patterns.

Overall, temporal continuity is strongly influenced by both the attribution mechanism and the evaluation dataset. CAM-based methods provide the most stable explanations across consecutive frames and exhibit the smallest degradation under cross-domain evaluation. Gradient-based methods offer limited temporal stability, with SmoothGrad representing a notable improvement through variance reduction. Perturbation-based methods are generally less temporally consistent, particularly KernelSHAP. These results highlight that temporal continuity is not guaranteed by high frame-level similarity alone, but requires stable localisation of salient regions across time, which is best achieved by activation-based attribution methods.

TABLE 6 | Temporal Continuity results across methods.

	FF-DF				Celeb-DF-v2				AVG			
Method	**Cos**	**IoU@10**	**IoU@20**	**IoU@30**	**Cos**	**IoU@10**	**IoU@20**	**IoU@30**	**Cos**	**IoU@10**	**IoU@20**	**IoU@30**
Score-CAM	0.97	0.81	0.86	0.88	0.92	0.65	0.73	0.76	0.95	0.73	0.79	0.82
CAM	0.97	0.83	0.87	0.86	0.95	0.77	0.80	0.78	0.96	0.80	0.83	0.82
Grad-CAM++	0.95	0.81	0.85	0.82	0.90	0.73	0.77	0.74	0.93	0.77	0.81	0.78
SmoothGrad	0.89	0.48	0.62	0.72	0.92	0.50	0.65	0.76	0.90	0.49	0.64	0.74
ExcitationBP	0.92	0.60	0.69	0.74	0.92	0.61	0.69	0.74	0.92	0.61	0.69	0.74
Grad-CAM	0.93	0.69	0.75	0.76	0.84	0.53	0.58	0.61	0.89	0.61	0.66	0.68
LRP	0.94	0.61	0.65	0.69	0.92	0.56	0.61	0.65	0.93	0.58	0.63	0.67
Occlusion Sensitivity	0.84	0.53	0.63	0.69	0.82	0.46	0.54	0.61	0.83	0.50	0.58	0.65
Vanilla gradient	0.68	0.31	0.50	0.62	0.67	0.30	0.46	0.58	0.68	0.30	0.48	0.60
Guided Backprop	0.65	0.36	0.48	0.59	0.62	0.34	0.48	0.58	0.63	0.35	0.48	0.58
DeepTaylor	0.95	0.42	0.48	0.54	0.95	0.39	0.45	0.51	0.95	0.40	0.47	0.52
Input × Gradient	0.63	0.29	0.43	0.53	0.62	0.26	0.39	0.49	0.62	0.28	0.41	0.51
Integrated gradients	0.63	0.28	0.40	0.50	0.62	0.27	0.39	0.48	0.62	0.27	0.40	0.49
Deconvolution	0.60	0.17	0.27	0.35	0.58	0.17	0.26	0.34	0.59	0.17	0.26	0.35
LIME	0.92	0.20	0.25	0.31	0.90	0.20	0.25	0.31	0.91	0.20	0.25	0.31
KernelSHAP	0.64	0.09	0.14	0.20	0.65	0.11	0.15	0.20	0.65	0.10	0.14	0.20

Note: The data highlighted in red represent the optimal value for each corresponding column. Cosine similarity measures frame-to-frame consistency of attribution maps. IoU@10, IoU@20 and IoU@30 denote the intersection-over-union between attribution maps of adjacent frames when retaining the top 10%, 20% and 30% most salient pixels, respectively (higher is better).

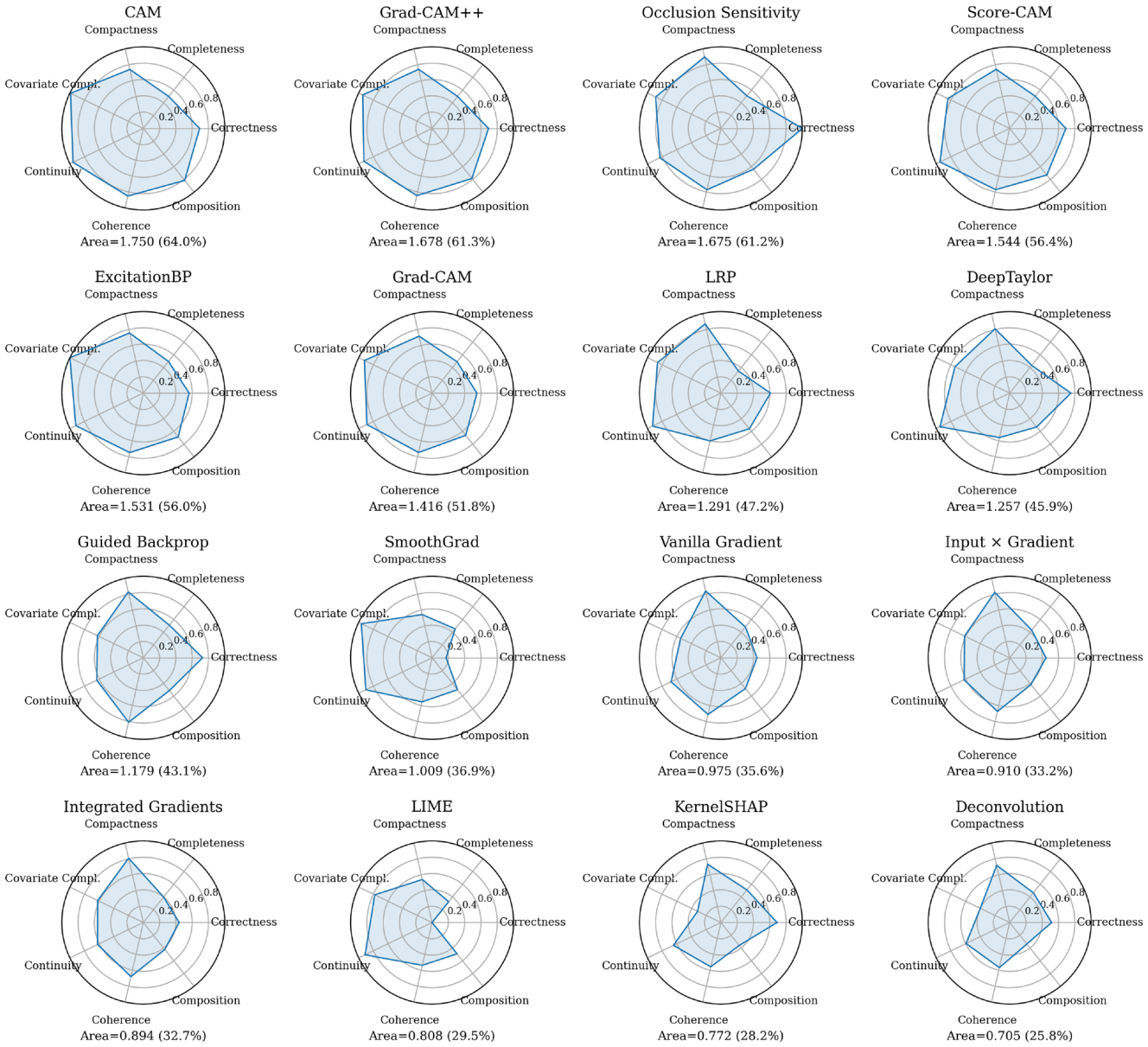

FIGURE 8 | Comparison of 16 explanation methods across seven Co-12 attributes.

4.2.5 | Overall Comparison

Figure 8 and Table 7 present a comparative analysis of 16 explanation methods evaluated across seven Co-12 attributes, each reflecting a distinct aspect of explanatory quality. The data highlighted in red represent the optimal value for each corresponding column. The radar charts encode normalised scores for each attribute:

- *Correctness*: Relative improvement in the PGI area under the curve, scaled to a unit interval.
- *Completeness*: Change in the model's confidence regarding a fake class at a predefined operating point
- *Compactness*: A geometric mean of the flip rate and inverse k-flip, promoting concise attributions.
- *Covariate Completeness*: Normalised adjusted F-ratio, reflecting class-separability in attribution maps.
- *Continuity*: Temporal consistency across frames, measured via mean cosine similarity.
- *Coherence*: Spatial alignment with ground-truth manipulated regions, quantified by intersection-over-union.
- *Composition*: A multiplicative measure integrating both spatial and temporal alignment.

The radar area of each method represents an unweighted aggregation over the seven Co-12 properties reported in Table 7, and serves as a compact summary of overall explanatory characteristics. The resulting ranking in terms of area and percentage reveals systematic differences in how explanation methods distribute performance across correctness, completeness, compactness, covariate completeness, temporal continuity, coherence and composition.

At the top of the ranking, *CAM* achieves the largest radar area (1.750, 64.0%), corresponding to consistently high values in

TABLE 7 | Co-12 properties scores of explainable methods.

Method	Correctness	Completeness	Compactness	Covariate comp.	Continuity	Coherence	Composition
CAM	0.23	0.50	0.19	30.69	0.96	0.85	0.82
Grad-CAM++	0.23	0.50	0.19	25.91	0.93	0.85	0.78
Score-CAM	0.23	0.50	0.19	17.91	0.95	0.77	0.73
ExcitationBP	0.19	0.50	0.19	31.64	0.92	0.75	0.69
Grad-CAM	0.18	0.49	0.18	23.83	0.89	0.75	0.66
Occlusion Sensitivity	0.33	0.51	0.33	21.17	0.83	0.77	0.64
LRP	0.20	0.35	0.28	19.22	0.93	0.60	0.56
DeepTaylor	0.25	0.43	0.22	12.44	0.95	0.56	0.53
Guided Backprop	0.24	0.51	0.23	7.84	0.63	0.81	0.51
SmoothGrad	0.06	0.45	0.13	27.73	0.90	0.55	0.50
LIME	−0.00	0.33	0.13	14.23	0.91	0.54	0.49
Vanilla Gradient	0.15	0.48	0.25	5.76	0.68	0.71	0.48
Integrated Gradients	0.15	0.40	0.23	7.72	0.62	0.68	0.42
Input × Gradient	0.15	0.43	0.23	7.50	0.62	0.67	0.42
KernelSHAP	0.23	0.51	0.18	2.02	0.65	0.56	0.36
Deconvolution	0.17	0.47	0.17	3.04	0.59	0.57	0.34

Note: The data highlighted in red represent the optimal value for each corresponding column.

temporal continuity (0.96), coherence (0.85) and composition (0.82), together with competitive correctness and completeness. *Grad-CAM++* follows closely (1.678, 61.3%), exhibiting a similarly balanced profile characterised by strong continuity (0.93), high coherence (0.85) and substantial covariate completeness (25.91). *Occlusion Sensitivity* ranks third overall (1.675, 61.2%) and is characterised by the highest correctness (0.33) and compactness (0.33) scores among all methods, while maintaining moderate temporal continuity (0.83), which contributes to a comparatively large and balanced radar polygon.

A second group of methods occupies the mid-to-high range of radar areas, including *Score-CAM* (1.544, 56.4%) and *ExcitationBP* (1.531, 56.0%). *Score-CAM* is associated with very high temporal continuity (0.95) and moderate coherence (0.77), whereas *ExcitationBP* attains the highest covariate completeness score (31.64), which substantially contributes to its overall area despite lower coherence and composition than *CAM*. *Grad-CAM* remains competitive (1.416, 51.8%), supported by strong continuity (0.89) and stable coherence (0.75), although comparatively lower correctness and compactness reduce its aggregated area.

Methods in the middle of the ranking exhibit more pronounced trade-offs across the evaluated properties. *LRP* (1.291, 47.2%) and *DeepTaylor* (1.257, 45.9%) are characterised by high temporal continuity (0.93–0.95), but lower coherence and composition constrain their overall radar areas. *Guided Backprop* (1.179, 43.1%) achieves the highest completeness score (0.51, tied for best) together with high coherence (0.81); however, its lower continuity (0.63) and limited covariate completeness (7.84) are associated with a reduced aggregated area, indicating weaker temporal stability across videos.

The lower-ranked methods generally exhibit unbalanced profiles or consistently lower values across multiple properties. *SmoothGrad* (1.009, 36.9%) shows strong continuity (0.90) and high covariate completeness (27.73), but its overall area is constrained by low correctness (0.06) and compactness (0.13). Simpler gradient-based variants, including *Vanilla Gradient* (0.975, 35.6%), *Input × Gradient* (0.910, 33.2%) and *Integrated Gradients* (0.894, 32.7%), achieve moderate completeness but are associated with weaker continuity and composition, resulting in smaller radar areas. Among perturbation-based methods, *LIME* (0.808, 29.5%) and *KernelSHAP* (0.772, 28.2%) are characterised by low composition and coherence together with limited covariate completeness, while *Deconvolution* yields the smallest radar area (0.705, 25.8%), reflecting comparatively low values across most evaluated dimensions.

Overall, the Co-12 score table and radar-area ranking indicate that no single explanation method achieves uniformly high performance across all properties. *CAM* and *Grad-CAM++* exhibit the most balanced profiles, driven by strong continuity, coherence and composition, whereas *Occlusion Sensitivity* represents a competitive alternative when correctness and compactness are emphasised. The remaining methods display application-dependent trade-offs, suggesting that the choice of explanation technique should be aligned with the specific Co-12 properties most relevant to the intended forensic scenario.

4.3 | Runtime Analysis

We conduct a comparative analysis of the computational efficiency of various explanation methods. The runtime is defined as the average time in seconds required to generate a single attribution map for one 256×256 face frame, under an identical hardware configuration. The results are summarised in Table 8.

Gradient-based approaches demonstrate superior computational efficiency. Specifically, the *CAM* method achieves the lowest average runtime of 0.030 s per frame, exhibiting a 2.4× speed-up relative to *Vanilla Gradient*, which records a runtime of 0.072 s. Other efficient methods include *Grad-CAM++* at 0.061 s and *Input×Gradient* at 0.064 s per frame.

In contrast, methods based on perturbation techniques incur significantly higher computational costs. *KernelSHAP* and *LIME* exhibit the highest runtimes, requiring 1.208 and 1.439 s per frame, respectively, corresponding to slowdowns of 26.7 and 20.0× relative to *Vanilla Gradient*. Intermediate runtime values are observed for methods such as *Integrated Gradients* (0.879 s), *SmoothGrad* (0.455 s) and *DeepTaylor* (0.444 s), reflecting their higher computational complexity.

These findings suggest that gradient-based explanation methods are more suitable for scenarios requiring efficient or real-time interpretability, while perturbation-based methods may be less appropriate in latency-sensitive contexts due to their substantial runtime overhead.

TABLE 8 | Runtime comparison of explanation methods.

Method	Avg. time per frame (s)	Relative speed
Vanilla Gradient	0.072	1.0×
Integrated Gradients	0.879	12.2× slower
SmoothGrad	0.455	6.3× slower
Input × Gradient	0.064	0.9× faster
Guided Backprop	0.084	1.2× slower
Deconvolution	0.084	1.2× slower
LRP	0.258	3.6× slower
DeepTaylor	0.444	6.2× slower
ExcitationBP	0.230	3.2× slower
Grad-CAM	0.105	1.5× slower
Grad-CAM++	0.061	1.2× faster
Score-CAM	0.191	2.7× slower
CAM	0.030	2.4× faster
Occlusion Sensitivity	0.172	2.6× slower
KernelSHAP	1.923	26.7× slower
LIME	1.439	20.0× slower

5 | Conclusion

This article has introduced a unified evaluation pipeline for post hoc visual explanation methods in deepfake detection, integrating four evaluation components that operationalise seven core XAI properties: *correctness*, *completeness*, *compactness*, *covariate completeness*, *continuity*, *composition* and *coherence*. Applying this pipeline to 16 representative methods revealed systematic differences across methodological families. CAM-based approaches demonstrated the most balanced performance across attributes, propagation-based methods achieved compact and stable attributions, gradient-based methods showed heterogeneous strengths and weaknesses, while perturbation-based methods performed comparatively poorly. Beyond benchmarking, the study contributes a reproducible pipeline that bridges conceptual interpretability frameworks and empirical practice, offering insights into method selection and forensic reliability. Nonetheless, the evaluation is limited by its focus on a single detector, as well as the exclusion of concept-based or inherently interpretable approaches. Future work should further extend the evaluation framework along several complementary dimensions. At the model level, adapting the pipeline to a wider range of detector architectures and training regimes would enable a more systematic analysis of generalisation behaviour and distributional shift across datasets. Beyond conventional accuracy-based assessment, robustness and faithfulness warrant deeper investigation through evaluation protocols that explicitly test the stability, determinism and evidential value of explanations under controlled changes in inputs, models or data conditions. In parallel, the conceptual structure and presentation of explanations merit closer examination, including the extent to which attributed features form compact, non-redundant and human-interpretable patterns. Another important direction concerns the incorporation of confidence-aware and knowledge-grounded assessments, so that explanation quality can also be judged in terms of plausibility and epistemic reliability. Finally, user-centred criteria remain essential: evaluating how explanations support realistic forensic tasks, how they align with practitioner needs and how interactive mechanisms may allow users to guide or refine explanatory outputs will help bridge the gap between technical evaluation and operational deployment. Advancing these lines of work will lead to a more complete and practically meaningful characterisation of explanation quality in deepfake detection systems.

Acknowledgements

Prof. Junxin Chen's work is supported by the Fundamental Research Funds for the Central Universities (No. DUT25YG229), and Xiaomi Young Talents Program. Prof. Salvador García's work is supported by the Spanish Ministry of Science and Technology under project PID2023-150070NB-I00. Prof. David Camacho's work is supported by the following projects: H2020 TMA-MSCA-DN TUAI project "Towards an Understanding of Artificial Intelligence via a transparent, open and explainable perspective" (HORIZON-MSCA-2023-DN-01-01, Grant agreement no: 101168344); by European Commission under IBERIFIER Plus—Iberian Digital Media Observatory (DIGITAL-2023-DEPLOY-04-EDMO-HUBS 101158511); and by Comunidad Autonoma de Madrid, CIRMA-CAM Project (TEC-2024/COM-404).

Funding

This work was supported by the Fundamental Research Funds for the Central Universities (No. DUT25YG229), and Xiaomi Young Talents Program. Prof. Salvador García's work is supported by the Spanish Ministry of Science and Technology under project PID2023-150070NB-I00. Prof. David Camacho's work is supported by the following projects: H2020 TMA-MSCA-DN TUAI project "Towards an Understanding of Artificial Intelligence via a transparent, open and explainable perspective" (HORIZON-MSCA-2023-DN-01-01, Grant agreement no: 101168344); by the European Commission under IBERIFIER Plus—Iberian Digital Media Observatory (DIGITAL-2023-DEPLOY-04-EDMO-HUBS 101158511); and by the Comunidad Autonoma de Madrid, CIRMA-CAM Project (TEC-2024/COM-404).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Adadi, A., and M. Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6: 52138–52160.

Adebayo, J., J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. 2018. "Sanity Checks for Saliency Maps." In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*, 9525–9536. Association for Computing Machinery.

Agarwal, C., S. Krishna, E. Saxena, et al. 2022. "OpenXAI: Towards a Transparent Evaluation of Model Explanations." NeurIPS Datasets and Benchmarks Track.

Aghasanli, A., D. Kangin, and P. Angelov. 2023. "Interpretable-Through-Prototypes Deepfake Detection for Diffusion Models." In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 467–474. IEEE.

Ai, G. 2019. Contributing Data to Deepfake Detection Research. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake.html.

Al-Ansari, N., D. Al-Thani, and R. S. Al-Mansoori. 2024. "User-Centered Evaluation of Explainable Artificial Intelligence (XAI): A Systematic Literature Review." *Human Behavior and Emerging Technologies* 2024, no. 1: 4628855.

Altmann, A., L. Toloşi, O. Sander, and T. Lengauer. 2010. "Permutation Importance: A Corrected Feature Importance Measure." *Bioinformatics* 26, no. 10: 1340–1347.

Alvarez-Melis, D., and T. Jaakkola. 2018. "Towards Robust Interpretability With Self-Explaining Neural Networks." *Advances in Neural Information Processing Systems* 31.

Anders, C. J., D. Neumann, W. Samek, K. R. Müller, and S. Lapuschkin. 2021. "Software for Dataset-Wide XAI: From Local Explanations to Global Insights With Zennit, CoRelAy, and ViRelAy." arXiv. abs/2106.13200.

Andreas, J., M. Rohrbach, T. Darrell, and D. Klein. 2016. "Neural Module Networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 39–48. IEEE.

Angelov, P. P., E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson. 2021. "Explainable Artificial Intelligence: An Analytical Review." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11, no. 5: e1424.

Anusha, T., and A. Srinagesh. 2025. "Deepfake Video Detection: A Comprehensive Survey of Advanced Machine Learning and Deep Learning Techniques to Combat Synthetic Video Manipulation." In *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*, 1033–1041. IEEE.

Bach, S., A. Binder, G. Montavon, F. Klauschen, K. R. Müller, and W. Samek. 2015. "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation." *PLoS One* 10, no. 7: e0130140.

Barredo Arrieta, A., N. Díaz-Rodríguez, J. Del Ser, et al. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI." *Information Fusion* 58: 82–115.

Belaid, M. K., R. Bornemann, M. Rabus, R. Krestel, and E. Hüllermeier. 2023. "Compare-xAI: Toward Unifying Functional Testing Methods for Post-Hoc XAI Algorithms Into a Multi-Dimensional Benchmark." In *Explainable Artificial Intelligence*, 88–109. Springer Nature Switzerland.

Bonomi, M., C. Pasquini, and G. Boato. 2020. "Dynamic Texture Analysis for Detecting Fake Faces in Video Sequences." arXiv:2007.15271. https://arxiv.org/abs/2007.15271.

Camilleri, M. A. 2024. "Artificial Intelligence Governance: Ethical Considerations and Implications for Social Responsibility." *Expert Systems* 41, no. 7: e13406.

Caruana, R., Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. 2015. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*, 1721–1730. Association for Computing Machinery.

Charte, D., F. Charte, S. García, d M J. Jesus, and F. Herrera. 2018. "A Practical Tutorial on Autoencoders for Nonlinear Feature Fusion: Taxonomy, Models, Software and Guidelines." *Information Fusion* 44: 78–96.

Chattopadhay, A., A. Sarkar, P. Howlader, and V. N. Balasubramanian. 2018. "Grad-Cam++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks." In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 839–847. IEEE.

Chen, C., O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. 2019. "This Looks Like That: Deep Learning for Interpretable Image Recognition." *Advances in Neural Information Processing Systems* 32: 8930–8941.

Chen, H., S. Lundberg, and S. I. Lee. 2019. "Explaining Models by Propagating Shapley Values of Local Components." arXiv:1911.11888. https://arxiv.org/abs/1911.11888.

Chen, J., Z. Guo, X. Xu, et al. 2024. "A Robust Deep Learning Framework Based on Spectrograms for Heart Sound Classification." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 21, no. 4: 936–947.

Chen, J., L. Song, M. Wainwright, and M. Jordan. 2018. "Learning to Explain: An Information-Theoretic Perspective on Model Interpretation." arXiv preprint arXiv:1802.07814.

Chen, J., W. Wang, B. Fang, et al. 2023. "Digital Twin Empowered Wireless Healthcare Monitoring for Smart Home." *IEEE Journal on Selected Areas in Communications* 41, no. 11: 3662–3676.

Chen, L., Y. Zhang, Y. Song, L. Liu, and J. Wang. 2022. "Self-Supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection." arXiv:2203.12208. https://arxiv.org/abs/2203.12208.

Chen, X., J. Chen, H. Xu, T. He, M. Fridenfalk, and Z. Lyu. 2025. "Smartphone-Based Measurement of Cardiovascular Healthcare: Advances and Applications." *Measurement* 252: 117347.

Chen, X., Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. 2016. "InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets." *Advances in Neural Information Processing Systems* 29: 2180–2188.

Cheng, L., Y. Liang, Y. Lu, and Y. M. Cheung. 2025. "GradToken: Decoupling Tokens With Class-Aware Gradient for Visual Explanation of Transformer Network." *Neural Networks* 181: 106837.

Chintha, A., B. Thai, S. J. Sohrawardi, et al. 2020. "Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection." *IEEE Journal of Selected Topics in Signal Processing* 14, no. 5: 1024–1037.

Chollet, F. 2017. "Xception: Deep Learning With Depthwise Separable Convolutions." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1800–1807. IEEE.

Chu, L., X. Hu, J. Hu, L. Wang, and J. Pei. 2019. Exact and Consistent Interpretation for Piecewise Linear Neural Networks: A Closed Form Solution. arXiv:1802.06259. https://arxiv.org/abs/1802.06259.

Ciftci, U. A., I. Demir, and L. Yin. 2024. "FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals." In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. IEEE.

Coroama, L., and A. Groza. 2022. "Evaluation Metrics in Explainable Artificial Intelligence (XAI)." In *Advanced Research in Technologies, Information, Innovation and Sustainability*, 401–413. Springer Nature Switzerland.

Cortés, D. G., E. Onieva, I. Pastor, L. Trinchera, and J. Wu. 2024. "Portfolio Construction Using Explainable Reinforcement Learning." *Expert Systems* 41, no. 11: e13667.

Dash, S., O. Günlük, and D. Wei. 2018. "Boolean Decision Rules via Column Generation." *Advances in Neural Information Processing Systems* 31: 4660–4670.

Dhurandhar, A., P. Y. Chen, R. Luss, et al. 2018. "Explanations Based on the Missing: Towards Contrastive Explanations With Pertinent Negatives." *Advances in Neural Information Processing Systems (NeurIPS)* 31: 11491.

Ding, X., Z. Raziei, E. C. Larson, E. V. Olinick, P. Krueger, and M. Hahsler. 2019. Swapped Face Detection Using Deep Learning and Subjective Assessment.

Dolhansky, B., J. Bitton, B. Pflaum, et al. 2020. "The DeepFake Detection Challenge (DFDC) Dataset." arXiv:2006.07397. https://arxiv.org/abs/2006.07397.

Doshi-Velez, F., and B. Kim. 2018. "Considerations for Evaluation and Generalization in Interpretable Machine Learning." In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, 3–17. Springer International Publishing.

Dwivedi, R., D. Dave, H. Naik, et al. 2023. "Explainable AI (XAI): Core Ideas, Techniques, and Solutions." *ACM Computing Surveys* 55, no. 9: 1–33.

Fang, K., J. Chen, H. Zhu, T. R. Gadekallu, X. Wu, and W. Wang. 2024. "Explainable-AI-Based Two-Stage Solution for WSN Object Localization Using Zero-Touch Mobile Transceivers." *Science China Information Sciences* 67, no. 7: 170302.

Friedman, J. H. 2001. "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29: 1189–1232.

Gao, J., Z. Xia, G. L. Marcialis, C. Dang, J. Dai, and X. Feng. 2024. "DeepFake Detection Based on High-Frequency Enhancement Network for Highly Compressed Content." *Expert Systems With Applications* 249: 123732.

Girón, A., J. Huertas-Tato, and D. Camacho. 2025. "Multimodal Visio-Lingual Content Analysis to Detect Fake Content on Reddit." In *Intelligent Data Engineering and Automated Learning – IDEAL 2024*, 143–154. Springer Nature Switzerland.

Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin. 2015. "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation." *Journal of Computational and Graphical Statistics* 24, no. 1: 44–65.

Gowrisankar, B., and V. L. Thing. 2024. "An Adversarial Attack Approach for eXplainable AI Evaluation on Deepfake Detection Models." *Computers & Security* 139: 103684.

Greenberg, S., and B. Buxton. 2008. "Usability Evaluation Considered Harmful (Some of the Time)." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*, 111–120. Association for Computing Machinery.

Guarnera, L., O. Giudice, and S. Battiato. 2020a. "Fighting Deepfake by Exposing the Convolutional Traces on Images." *IEEE Access* 8: 165085–165098.

Guarnera, L., O. Giudice, and S. Battiato. 2020b. "DeepFake Detection by Analyzing Convolutional Traces." arXiv:2004.10448. https://arxiv.org/abs/2004.10448.

Gunning, D., and D. Aha. 2019. "DARPA's Explainable Artificial Intelligence (XAI) Program." *AI Magazine* 40, no. 2: 44–58.

Guo, S., M. Gao, Q. Li, G. Jeon, and D. Camacho. 2024. "Deepfake Detection via a Progressive Attention Network." In *2024 International Joint Conference on Neural Networks (IJCNN)*, 1–6. IEEE.

Gurumoorthy, K. S., A. Dhurandhar, G. Cecchi, and C. Aggarwal. 2019. "Efficient Data Representation by Selecting Prototypes With Importance Weights." In *2019 IEEE International Conference on Data Mining (ICDM)*, 260–269. IEEE.

Habbal, A., M. K. Ali, and M. A. Abuzaraida. 2024. "Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, Applications, Challenges and Future Research Directions." *Expert Systems With Applications* 240: 122442.

Haq, I. U., K. M. Malik, and K. Muhammad. 2024. "Multimodal Neurosymbolic Approach for Explainable Deepfake Detection." *ACM Transactions on Multimedia Computing, Communications, and Applications* 20, no. 11: 1–16.

Heidari, A., N. Jafari Navimipour, H. Dag, and M. Unal. 2024. "Deepfake Detection Using Deep Learning Methods: A Systematic and Comprehensive Review." *WIREs Data Mining and Knowledge Discovery* 14, no. 2: e1520.

Higgins, I., L. Matthey, A. Pal, et al. 2017. "Beta-VAE: Learning Basic Visual Concepts With a Constrained Variational Framework." In *Proceedings of the International Conference on Learning Representations*. IEEE.

Holzinger, A., A. Saranti, C. Molnar, P. Biecek, and W. Samek. 2022. "Explainable AI Methods - A Brief Overview." In *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction With ICML 2020, Vienna, Austria, Revised and Extended Papers*, 13–38. Springer International Publishing.

Huang, X., R. Li, Y. M. Cheung, K. C. Cheung, S. See, and R. Wan. 2024. "GaussianMarker: Uncertainty-Aware Copyright Protection of 3D Gaussian Splatting" arXiv:2410.23718. https://arxiv.org/abs/2410.23718.

Huang, Z., and Y. Li. 2020. "Interpretable and Accurate Fine-Grained Recognition via Region Grouping." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8662–8672. IEEE.

Jacovi, A., and Y. Goldberg. 2020. "Towards Faithfully Interpretable NLP Systems: How Should we Define and Evaluate Faithfulness?" In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 4198–4205. Association for Computing Machinery.

Jafar, M. T., M. Ababneh, M. Al-Zoube, and A. Elhassan. 2020. "Forensics and Analysis of Deepfake Videos." In *2020 11th International Conference on Information and Communication Systems (ICICS)*, 53–58. IEEE.

Jayakumar, K., and N. Skandhakumar. 2022. "A Visually Interpretable Forensic Deepfake Detection Tool Using Anchors." In *2022 7th International Conference on Information Technology Research (ICITR)*, 1–6. IEEE.

Jiang, L., R. Li, W. Wu, C. Qian, and C. C. Loy. 2020. "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection." arXiv preprint arXiv:2001.03024.

Júnior, J. S. S., C. Gaspar, J. Mendes, and C. Premebida. 2024. "Assessing Interpretability of Data-Driven Fuzzy Models: Application in Industrial Regression Problems." *Expert Systems* 41, no. 12: e13710.

Keita, M., W. Hamidouche, H. Bougueffa Eutamene, A. Taleb-Ahmed, D. Camacho, and A. Hadid. 2025. "Bi-LORA: A Vision-Language Approach for Synthetic Image Detection." *Expert Systems* 42, no. 2: e13829.

Khalaf, O., A. B. Ishak, and S. García. 2026. "Towards Explainable Multilabel Learning: Fusing Label Dependency Analysis With Monotonic Decision Trees." *Information Fusion* 127: 103691.

Khan, M. M. 2023. "Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods." *Intelligent Automation & Soft Computing* 35: 2151–2169.

Khodabakhsh, A., R. Ramachandra, K. Raja, P. Wasnik, and C. Busch. 2018. "Fake Face Detection Methods: Can They be Generalized?" In *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 1–6. IEEE.

Kim, B., M. Wattenberg, J. Gilmer, et al. 2017. "Interpretability Beyond Feature Attribution: Quantitative Testing With Concept Activation Vectors (TCAV)." In *Proceedings of the 34th International Conference on Machine Learning*. PLMR.

Kim, B., M. Wattenberg, J. Gilmer, et al. 2018. "Interpretability Beyond Feature Attribution: Quantitative Testing With Concept Activation Vectors (Tcav)." In *Proceedings of the International Conference on Machine Learning (ICML)*, 2668–2677. PLMR.

Kindermans, P. J., K. T. Schütt, M. Alber, et al. 2017. "Learning How to Explain Neural Networks: Patternnet and Patternattribution." arXiv preprint arXiv:1705.05598.

Koh, P. W., T. Nguyen, Y. S. Tang, et al. 2020. "Concept Bottleneck Models." In *Proceedings of the 37th International Conference on Machine Learning*, 5338–5348. Association for Computing Machinery.

Kokhlikyan, N., V. Miglani, M. Martin, et al. 2020. "Captum: A Unified and Generic Model Interpretability Library for PyTorch." arXiv:2009.07896. http://arxiv.org/abs/2009.07896.

Korshunov, P., and S. Marcel. 2018. "DeepFakes: A New Threat to Face Recognition? Assessment and Detection." arXiv preprint arXiv:1812.08685.

Lage, I., E. Chen, J. He, et al. 2019. "An Evaluation of the Human-Interpretability of Explanation." arXiv Preprint arXiv:1902.00006.

Lakkaraju, H., and J. Leskovec. 2016. "Confusions Over Time: An Interpretable Bayesian Model to Characterize Trends in Decision Making." *Advances in Neural Information Processing Systems (NeurIPS)* 29: 3269–3277.

Leavitt, M. L., and A. Morcos. 2020. "Towards Falsifiable Interpretability Research." arXiv:2010.12016. https://arxiv.org/abs/2010.12016.

Li, J., L. Liu, H. Li, G. Li, G. Huang, and S. Shi. 2020. "Evaluating Explanation Methods for Neural Machine Translation" arXiv:2005.01672. https://arxiv.org/abs/2005.01672.

Li, Q., M. Gao, G. Zhang, W. Zhai, J. Chen, and G. Jeon. 2024. "Towards Multimodal Disinformation Detection by Vision-Language Knowledge Interaction." *Information Fusion* 102: 102037.

Li, Y., X. Yang, P. Sun, H. Qi, and S. Lyu. 2020. "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics." In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3204–3213. IEEE.

Linardatos, P., V. Papastefanopoulos, and S. Kotsiantis. 2021. "Explainable AI: A Review of Machine Learning Interpretability Methods." *Entropy* 23, no. 1: 18.

Liu, H., X. Li, W. Zhou, et al. 2021. "Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain." arXiv:2103.01856. https://arxiv.org/abs/2103.01856.

Liz-López, H., M. Keita, A. Taleb-Ahmed, A. Hadid, J. Huertas-Tato, and D. Camacho. 2024. "Generation and Detection of Manipulated Multimodal Audiovisual Content: Advances, Trends and Open Challenges." *Information Fusion* 103, no. C: 102103.

Lou, Y., R. Caruana, J. Gehrke, and G. Hooker. 2013. "Accurate Intelligible Models With Pairwise Interactions." In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 623–631. Association for Computing Machinery.

Lundberg, S. M., and S. I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems (NeurIPS)* 30: 4768–4777.

Luo, Y., Y. Zhang, J. Yan, and W. Liu. 2021. "Generalizing Face Forgery Detection With High-Frequency Features." arXiv:2103.12376. https://arxiv.org/abs/2103.12376.

Malolan, B., A. Parekh, and F. Kazi. 2020. "Explainable Deep-Fake Detection Using Visual Interpretability Methods." In *2020 3rd International Conference on Information and Computer Technologies (ICICT)*, 289–293. IEEE.

Mansoor, N., and A. I. Iliev. 2025. "Explainable AI for DeepFake Detection." *Applied Sciences* 15, no. 2: 725.

Masi, I., A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed. 2020. "Two-Branch Recurrent Network for Isolating Deepfakes in Videos." arXiv:2008.03412. https://arxiv.org/abs/2008.03412.

Matern, F., C. Riess, and M. Stamminger. 2019. "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations." In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 83–92. IEEE.

McCullagh, P. 2019. *Generalized Linear Models*. Routledge.

Nam, W. J., S. Gur, J. Choi, L. Wolf, and S. W. Lee. 2020. "Relative Attributing Propagation: Interpreting the Comparative Contributions of Individual Units in Deep Neural Networks." *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* 34, no. 3: 2501–2508.

Nauta, M., J. Trienes, S. Pathak, et al. 2023. "From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI." *ACM Computing Surveys* 55, no. 13s: 1–42.

Nguyen, H. H., J. Yamagishi, and I. Echizen. 2019. "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos." In *ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2307–2311. IEEE.

Nie, J., Y. Zhang, T. Liu, Y. M. Cheung, B. Han, and X. Tian. 2024. "Detecting Discrepancies Between AI-Generated and Natural Images Using Uncertainty." arXiv:2412.05897. https://arxiv.org/abs/2412.05897.

Oztireli, A. C., M. Ancona, E. Ceolini, and M. Gross. 2019. "Towards Better Understanding of Gradient-Based Attribution Methods for Deep Neural Networks." ArXiv Preprint.

Pan, D., L. Sun, R. Wang, X. Zhang, and R. O. Sinnott. 2020. "Deepfake Detection Through Deep Learning." In *2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, 134–143. IEEE.

Passos, L. A., D. Jodas, K. A. P. Costa, et al. 2024. "A Review of Deep Learning-Based Approaches for Deepfake Content Detection." *Expert Systems* 41, no. 8: e13570.

Petsiuk, V., A. Das, and K. Saenko. 2018. "Rise: Randomized Input Sampling for Explanation of Black-Box Models." arXiv preprint arXiv:1806.07421.

Plumb, G., M. Al-Shedivat, A. A. Cabrera, A. Perer, E. Xing, and A. Talwalkar. 2020. Regularizing Black-Box Models for Improved Interpretability. arXiv:1902.06787. https://arxiv.org/abs/1902.06787.

Qureshi, K. N., H. Nafea, and P. Kim. 2025. "Advancing Healthcare Systems: A Tri-Tier Architecture by Using Data Communication, AI Data Generative and Regulation and Compliance Standards." *Expert Systems* 42, no. 2: e13742. https://doi.org/10.1111/exsy.13742.

Rana, M. S., M. N. Nobi, B. Murali, and A. H. Sung. 2022. "Deepfake Detection: A Systematic Literature Review." *IEEE Access* 10: 25494–25513.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. "Why Should I Trust You? Explaining the Predictions of Any Classifier." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, 1135–1144. Association for Computing Machinery.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2018. "Anchors: High-Precision Model-Agnostic Explanations." *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* 32, no. 1: 1527–1535.

Robnik-Šikonja, M., and M. Bohanec. 2018. "Perturbation-Based Explanations of Prediction Models." In *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, 159–175. Springer International Publishing.

Ross, A. S., M. C. Hughes, and F. Doshi-Velez. 2017. "Right for the Right Reasons: Training Differentiable Models by Constraining Their Explanations." arXiv preprint arXiv:1703.03717.

Rössler, A., D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner. 2019. "FaceForensics++: Learning to Detect Manipulated Facial Images." In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1–11. IEEE.

Rössler, A., D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. 2018. FaceForensics: A Large-Scale Video Dataset for Forgery Detection in Human Faces. arXiv preprint arXiv:1803.09179.

Rosso, P., B. Chulvi, D. Korenčić, et al. 2024. XAI-DisInfodemics: eXplainable AI for Disinformation and Conspiracy Detection During Infodemics. *SEPLN-CEDI-PD* 2024: *Seminar of the Spanish Society for Natural Language Processing: Projects and System Demonstrations (CEUR Workshop Proceedings)*.

Sabir, E., J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. 2019. "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos." arXiv:1905.00582. https://arxiv.org/abs/1905.00582.

Sandhya, and A. A. Kashyap. 2025. "A Statistical Analysis for Deepfake Videos Forgery Traces Recognition Followed by a Fine-Tuned InceptionResNetV2 Detection Technique." *Journal of Forensic Sciences* 70, no. 1: 349–368.

Schwab, P., and W. Karlen. 2019. "CXPlain: Causal Explanations for Model Interpretation Under Uncertainty." Advances in Neural Information Processing Systems (NeurIPS).

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. "Grad-Cam: Visual Explanations From Deep Networks via Gradient-Based Localization." In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626. IEEE.

Shi, W., H. Zhou, N. Miao, and L. Li. 2020. Dispersed Exponential Family Mixture VAEs for Interpretable Text Generation. arXiv:1906.06719. https://arxiv.org/abs/1906.06719.

Shrikumar, A., P. Greenside, and A. Kundaje. 2017. "Learning Important Features Through Propagating Activation Differences." In *Proceedings of the International Conference on Machine Learning (ICML)*, 3145–3153. PLMR.

Silva, S. H., M. Bethany, A. M. Votto, I. H. Scarff, N. Beebe, and P. Najafirad. 2022. "Deepfake Forensics Analysis: An Explainable Hierarchical Ensemble of Weakly Supervised Models." *Forensic Science International: Synergy* 4: 100217.

Simonyan, K., A. Vedaldi, and A. Zisserman. 2013. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." arXiv preprint arXiv:1312.6034.

Singh, A., R. K. Shrivastava, and A. Srivastava. 2025. "Efficient and Compressed Deep Learning Model for Brain Tumour Classification With Explainable AI for Smart Healthcare and Information Communication Systems." *Expert Systems* 42, no. 2: e13770.

Smilkov, D., N. Thorat, B. Kim, F. Viégas, and M. Wattenberg. 2017. "Smoothgrad: Removing Noise by Adding Noise." arXiv preprint arXiv:1706.03825.

Springenberg, J. T., A. Dosovitskiy, T. Brox, and M. Riedmiller. 2014. "Striving for Simplicity: The All Convolutional Net." arXiv preprint arXiv:1412.6806.

Sudarshana, K., and Y. Vamsidhar. 2025. "UAM-Net: Robust Deepfake Detection Through Hybrid Attention Into Scalable Convolutional Network." *Expert Systems* 42, no. 3: e70009.

Sundararajan, M., A. Taly, and Q. Yan. 2017. "Axiomatic Attribution for Deep Networks." In *Proceedings of the International Conference on Machine Learning (ICML)*, 3319–3328. PLMR.

Tariq, S., S. Lee, and S. S. Woo. 2020. "A Convolutional LSTM Based Residual Network for Deepfake Video Detection." arXiv:2009.07480. https://arxiv.org/abs/2009.07480.

The Royal Society. 2019. Explainable AI: The Basics. Royal Society Policy Briefing, Policy Briefing (online).

Tsigos, K., E. Apostolidis, S. Baxevanakis, S. Papadopoulos, and V. Mezaris. 2024. "Towards Quantitative Evaluation of Explainable AI Methods for Deepfake Detection." In *Proceedings of the 3rd ACM International Workshop on Multimedia AI Against Disinformation (MAD '24)*, 37–45. Association for Computing Machinery.

Ustun, B., and C. Rudin. 2016. "Supersparse Linear Integer Models for Optimized Medical Scoring Systems." *Machine Learning* 102, no. 3: 349–391.

Wani, N. A., R. Kumar, Mamta, J. Bedi, and I. Rida. 2024. "Explainable AI-Driven IoMT Fusion: Unravelling Techniques, Opportunities, and Challenges With Explainable AI in Healthcare." *Information Fusion* 110: 102472.

Xu, F., H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu. 2019. "Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges." In *Natural Language Processing and Chinese Computing*, 563–574. Springer International Publishing.

Xu, X., J. Chen, W. Lv, W. Wang, and Y. Zhang. 2025. "Image Tampering Detection With Frequency-Aware Attention and Multiview Fusion." *IEEE Transactions on Artificial Intelligence* 6, no. 3: 614–625.

Xu, X., W. Lv, W. Wang, Y. Zhang, and J. Chen. 2024. "Empowering Semantic Segmentation With Selective Frequency Enhancement and Attention Mechanism for Tampering Detection." *IEEE Transactions on Artificial Intelligence* 5, no. 6: 3270–3283.

Yan, Z., T. Yao, S. Chen, et al. 2024. "DF40: Toward Next-Generation Deepfake Detection." *Advances in Neural Information Processing Systems* 925: 29387–29434.

Yan, Z., Y. Zhang, X. Yuan, S. Lyu, and B. Wu. 2023. "DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection." *Advances in Neural Information Processing Systems (NeurIPS)* 36: 4534–4565.

Yang, X., Y. Li, and S. Lyu. 2018. "Exposing Deep Fakes Using Inconsistent Head Poses." arXiv:1811.00661. https://arxiv.org/abs/1811.00661.

Yang, X., Y. Li, and S. Lyu. 2019. "Exposing Deep Fakes Using Inconsistent Head Poses." In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 8261–8265. IEEE.

Yazdinejad, A., R. M. Parizi, G. Srivastava, and A. Dehghantanha. 2020. "Making Sense of Blockchain for AI Deepfakes Technology." In *2020 IEEE Globecom Workshops (GC Wkshps)*, 1–6. IEEE.

Yoon, J., A. Panizo-LLedot, D. Camacho, and C. Choi. 2024. "Triple-Modality Interaction for Deepfake Detection on Zero-Shot Identity." *Information Fusion* 109, no. C: 102424.

Zhao, T., X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia. 2021. "Learning Self-Consistency for Deepfake Detection." arXiv:2012.09311. https://arxiv.org/abs/2012.09311.

Zhou, B., A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. 2016. "Learning Deep Features for Discriminative Localization." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2921–2929. IEEE.

Zhou, P., X. Han, V. I. Morariu, and L. S. Davis. 2017. "Two-Stream Neural Networks for Tampered Face Detection." In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 1831–1839. IEEE.

Zi, B., M. Chang, J. Chen, X. Ma, and Y. G. Jiang. 2020. "WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection." In *Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery.

Appendix A

Algorithmic Details

ALGORITHM A1 | Vanilla gradient.

Require: detector f, batch dict D with $x = D[\texttt{image}] \in \mathbb{R}^{B\times 3\times H\times W}$

$x \leftarrow \texttt{clone\_detached}(D[\texttt{image}]);\ \texttt{requires\_grad}(x) \leftarrow \texttt{True}$

$D' \leftarrow \texttt{copy}(D);\ D'[\texttt{image}] \leftarrow x$

$y \leftarrow f(D', \texttt{inference} = \texttt{True});\ \phi \leftarrow y[\texttt{prob}] \in \mathbb{R}^{B}$

$f.\texttt{zero\_grad}();\ \texttt{backward}(\sum_b \phi_b)$

$G \leftarrow \ |\partial(\sum_b \phi_b)/\partial x| \in \mathbb{R}^{B\times 3\times H\times W}$

$\texttt{Parent.SaveExplanation}(G, D')$; **return** G.

ALGORITHM A2 | LRP.

Require: Model f, input batch $x \in \mathbb{R}^{B\times 3\times H\times W}$, composite $C = \texttt{EpsilonPlus}$
Ensure: Attribution $A^{\uparrow} \in \mathbb{R}^{B\times 3\times 256\times 256}$

Enable gradients on x; $y \leftarrow f(x)$
if $y \in \mathbb{R}^{B\times C}$ **then**
    $t_b \leftarrow \arg\max_k y_b^{(k)}$; $\phi(y)_b \leftarrow e_{t_b} \quad \forall b$
else
    $\phi(y)_b \leftarrow 1 \quad \forall b$
end if
G $\leftarrow$ ZGradient(f, C); $(\_, A) \leftarrow \mathrm{G}(x, \phi)$
if $(H, W) \neq (256,256)$ **then**
    $A^{\uparrow} \leftarrow \mathrm{Interp}_{256\times 256}(A;\ \texttt{bilinear},\ \texttt{align\_corners} = \texttt{False})$
else
    $A^{\uparrow} \leftarrow A$
end if
return $A^{\uparrow}$.

ALGORITHM A3 | DeepTaylor.

Require: Model f, input batch $x \in \mathbb{R}^{B\times C\times H\times W}$, Box composite $C = \texttt{Box}(\ell, h)$
Ensure: Attribution $A^{\uparrow} \in \mathbb{R}^{B\times 3\times 256\times 256}$

Enable gradients on x; $y \leftarrow f(x)$
if $y \in \mathbb{R}^{B\times K}$ **then**
  $t_b \leftarrow \arg\max_k y_b^{(k)}$; $\phi(y)_b \leftarrow e_{t_b} \quad \forall b$
else
  $\phi(y)_b \leftarrow 1 \quad \forall b$
end if
Infer bounds (ℓ, h) **in input space, with shape** $B \times C \times H \times W$:
if x already in $[0, 1]$ **then**
  $\ell \leftarrow 0, \ h \leftarrow 1$
else
  $\mu = (0.485, 0.456, 0.406), \ \sigma = (0.229, 0.224, 0.225)$ (per-channel)
  $\ell_{:,c,:,:} \leftarrow \left(0 - \mu_c\right) / \sigma_c, \quad h_{:,c,:,:} \leftarrow \left(1 - \mu_c\right) / \sigma_c$ (broadcast to $B \times C \times H \times W$)
end if
$C \leftarrow \texttt{Box}(\ell, h)$
G $\leftarrow$ ZGradient(f, C)
$(\_, A) \leftarrow$ G(x, ϕ) ▷ $A \in \mathbb{R}^{B\times 3\times H\times W}$
if $(H, W) \neq (256, 256)$ **then**
  $A^{\uparrow} \leftarrow \text{Interp}_{256\times 256}(A;\ \texttt{bilinear},\ \texttt{align\_corners} = \texttt{False})$
else.
  $A^{\uparrow} \leftarrow A$
end if
return $A^{\uparrow}$

ALGORITHM A4 | ExcitationBP.

Require: Model f, input batch $x \in \mathbb{R}^{B\times 3\times H\times W}$, composite $C = \texttt{ExcitationBP}$
Ensure: Attribution $A^{\uparrow} \in \mathbb{R}^{B\times 3\times 256\times 256}$

Enable gradients on x; $y \leftarrow f(x)$
if $y \in \mathbb{R}^{B\times C}$ **then**
  $t_b \leftarrow \arg\max_k y_b^{(k)}$; $\phi(y)_b \leftarrow e_{t_b} \quad \forall b$
else
  $\phi(y)_b \leftarrow 1 \quad \forall b$
end if
G $\leftarrow$ ZGradient(f, C); $(\_, A) \leftarrow$ G(x, ϕ)
if $(H, W) \neq (256, 256)$ **then**
  $A^{\uparrow} \leftarrow \text{Interp}_{256\times 256}(A;\ \texttt{bilinear},\ \texttt{align\_corners} = \texttt{False})$
else
  $A^{\uparrow} \leftarrow A$
end if
return $A^{\uparrow}$

ALGORITHM A5 | Grad-CAM.

Require: Model f, input batch $x \in \mathbb{R}^{B\times 3\times H\times W}$, auto-picked target layer

Ensure: Attribution $A^{\uparrow} \in \mathbb{R}^{B\times 3\times 256\times 256}$

Enable gradients on x; register forward/backward hooks at target layer

$p \leftarrow f(\{\text{image}: x\}, \texttt{inference} = \texttt{True})[\texttt{prob}]$

$J(x) \leftarrow \sum_{b=1}^{B} p_b$; **backward** from $J(x)$ to capture $A \in \mathbb{R}^{B\times C\times h\times w}$ and $g = \partial J / \partial A$

$w_k \leftarrow \frac{1}{hw} \sum_{i,j} g_{kij}$

$\text{cam} \leftarrow \text{ReLU}\left(\sum_k w_k A_k \right)$

$\text{cam} \leftarrow \text{Interp}_{H\times W}(\text{cam}; \texttt{bilinear}, \texttt{align\_corners} = \texttt{False})$

$A \leftarrow \text{RepeatChannel}(\text{cam}, 3)$

$A^{\uparrow} \leftarrow A$

return $A^{\uparrow}$

ALGORITHM A6 | Grad-CAM++.

Require: Model f, input batch $x \in \mathbb{R}^{B\times 3\times H\times W}$, auto-picked target layer

Ensure: Attribution $A^{\uparrow} \in \mathbb{R}^{B\times 3\times 256\times 256}$

Enable gradients on x; register hooks; $p \leftarrow f(\{\text{image}: x\}, \texttt{inference} = \texttt{True})[\texttt{prob}]$

$J(x) \leftarrow \sum_{b=1}^{B} p_b$; **backward** to capture A and $g = \partial J / \partial A$; set $\varepsilon = 10^{-8}$.

$g^2 \leftarrow g \odot g$; $g^3 \leftarrow g^2 \odot g$; $S \leftarrow \sum_{i,j} A \odot g^3$

$\alpha \leftarrow \max\left(0, \frac{g^2}{2g^2 + S + \varepsilon}\right)$

$w_k \leftarrow \sum_{i,j} \alpha_{kij} \text{ReLU}(g_{kij})$

$\text{cam} \leftarrow \text{ReLU}\left(\sum_k w_k A_k \right)$;

$A \leftarrow \text{RepeatChannel}(\text{cam}, 3)$; resize to 256×256 if needed.

return $A^{\uparrow}$

ALGORITHM A7 | Score-CAM.

Require: Model f, input batch $x \in \mathbb{R}^{B\times 3\times H\times W}$, auto-picked target layer, top-K channels, mask batch size b_s

Ensure: Attribution $A^{\uparrow} \in \mathbb{R}^{B\times 3\times 256\times 256}$

Register forward hook; run f once to cache $A \in \mathbb{R}^{B\times C\times h\times w}$

Rank channels by $\text{mean}_{b,i,j}(A_{kij})$; keep indices $\mathcal{K}$ of top-K

for each $k \in \mathcal{K}$ **do**

    $M_k \leftarrow \text{norm}\left(\uparrow \text{ReLU}(A_k) \right)$

end for

for $b = 1$ **to** B **do**

    for masks in chunks of size b_s **do**

        $x_{\text{masked}} \leftarrow x^{(b)} \odot \left\{ M_k^{(b)} \right\}$; $p \leftarrow f\left(\{x_{\text{masked}}\}, \texttt{inference} = \texttt{True} \right)[\texttt{prob}]$

        Assign weights $w_k^{(b)} \leftarrow p$ for the corresponding masks

    end for

end for

$\text{cam}^{(b)} \leftarrow \text{ReLU}\left(\sum_{k\in\mathcal{K}} w_k^{(b)} M_k^{(b)} \right)$ (already at (H, W))

$A \leftarrow \text{RepeatChannel}(\text{cam}, 3)$; $A^{\uparrow} \leftarrow$ resize to 256×256 if needed

return $A^{\uparrow}$.

ALGORITHM A8 | Classic CAM.

Require: Model f, input batch $x \in \mathbb{R}^{B\times 3\times H\times W}$, target layer, linear classifier weights $W \in \mathbb{R}^{K\times C}$
Ensure: Attribution $A^{\uparrow} \in \mathbb{R}^{B\times 3\times 256\times 256}$

Register forward hook; $p \leftarrow f(\{\text{image: } x\}, \texttt{inference} = \texttt{True})[\texttt{prob}]$ ▷ to cache A
Obtain $A \in \mathbb{R}^{B\times C\times h\times w}$ from hook; set $\widetilde{w}_c \leftarrow \sum_{k=1}^{K} W_{kc}$
cam $\leftarrow \mathrm{ReLU}\left(\sum_{c=1}^{C} \widetilde{w}_c A_c\right)$; upsample to (H, W)
$A \leftarrow \mathrm{RepeatChannel}(\text{cam}, 3)$; $A^{\uparrow} \leftarrow$ resize to 256×256 if needed
return $A^{\uparrow}$

ALGORITHM A9 | Occlusion sensitivity.

Require: Model f, input batch $x \in \mathbb{R}^{B\times 3\times H\times W}$, window (h_w, w_w), stride (s_h, s_w), baseline mode
Ensure: Attributions $A \in \mathbb{R}^{B\times 3\times H\times W}$

Define $\texttt{forward\_func}(\cdot) \leftarrow f(\{\text{image: } \cdot\}, \texttt{inference} = \texttt{True})[\texttt{prob}]$
Occ $\leftarrow \texttt{Occlusion}(\texttt{forward\_func})$; $A \leftarrow \emptyset$
for $i = 1$ **to** B **do**.
    $x_i \leftarrow x[i]$; $b \leftarrow \begin{cases} \mathrm{mean}(x_i), & \text{use-image-mean} \\ 0, & \text{otherwise} \end{cases}$
    $\widetilde{A} \leftarrow \text{Occ.}\texttt{attribute}(x_i,\ \texttt{sliding\_window\_shapes} = (3, h_w, w_w),\ \texttt{strides} = (3, s_h, s_w),\ \texttt{baselines} = b)$.
    $\widetilde{A} \leftarrow |\widetilde{A}|$ ▷ match gradient-style magnitude
    Append $\widetilde{A}$ to A
end for
return A

ALGORITHM A10 | KernelSHAP attribution.

Require: Model f, input batch $x \in \mathbb{R}^{B\times 3\times H\times W}$, patch size p, n_{samples}, `perturbations_per_eval`, baseline mode
Ensure: Attributions $A \in \mathbb{R}^{B\times 3\times H\times W}$

Define $\texttt{forward\_func}(\cdot) \leftarrow f(\{\text{image: } \cdot\}, \texttt{inference} = \texttt{True})[\texttt{prob}]$
KS $\leftarrow \texttt{KernelShap}(\texttt{forward\_func})$; $A \leftarrow \emptyset$
Build feature mask $M \in \mathbb{N}^{3\times H\times W}$ by tiling $p \times p$ patches across (H, W) with shared IDs across channels
for $i = 1$ **to** B **do**
    $x_i \leftarrow x[i]$; $b \leftarrow \begin{cases} \mathrm{mean}(x_i), & \texttt{baseline\_mode} = \texttt{mean} \\ 0, & \texttt{zero} \\ \texttt{baseline\_const}, & \texttt{const} \end{cases}$
    $\widetilde{A} \leftarrow \text{KS.}\texttt{attribute}(x_i,\ \texttt{baselines} = b,\ \texttt{feature\_mask} = M,\ n_{\text{samples}},\ \texttt{perturbations\_per\_eval},\ \texttt{show\_progress} = \texttt{False})$
    $\widetilde{A} \leftarrow |\widetilde{A}|$ ▷ absolute attribution per patch broadcast to pixels
    Append $\widetilde{A}$ to A
end for
return A

ALGORITHM A11 | LIME (image).

Require: Model f, input batch $x \in \mathbb{R}^{B\times 3\times H\times W}$, `num_samples`, `batch_size`, target label $\ell = 1$, optional resize (W', H'), SLIC params $(n_{\text{seg}},\ \text{compactness},\ \sigma)$

Ensure: Attributions $A \in \mathbb{R}^{B\times 3\times H\times W}$

L $\leftarrow$ `LimeImageExplainer(random_state)`; $A \leftarrow \emptyset$

Define classifier callback C: stack input images $\rightarrow$ tensor $\rightarrow$ p_{fake} = `forward_prob`; return $[p_{\text{real}}, p_{\text{fake}}]$

for $i = 1$ **to** B **do**

  $x_i \leftarrow x[i]$; convert to $I \in \mathbb{R}^{H\times W\times 3}$

  $I' \leftarrow \begin{cases} \texttt{resize}(I \rightarrow W'\times H'), & \text{if specified} \\ I, & \text{otherwise} \end{cases}$; $c_{\text{hide}} \leftarrow \text{mean}(I')\,(\text{or } 0)$

  `segments` $\leftarrow$ `SLIC`$(I';\ n_{\text{seg}},\ \text{compactness},\ \sigma,\ \texttt{start\_label} = 0)$

  `exp` $\leftarrow$ L.`explain_instance`$(I', C, \texttt{labels} = (\ell), \texttt{hide\_color} = c_{\text{hide}},$ `num_samples, batch_size, segmentation_fn` = `SLIC`)

  Build low-res weight map W' by assigning each superpixel its weight from `exp.local_exp`$[\ell]$

  $W \leftarrow$ `bilinear_resize`$(W' \rightarrow W \times H)$ if needed; $W \leftarrow (W - \min W)/(\max W - \min W + \varepsilon)$

  $\widetilde{A} \leftarrow$ `RepeatChannel`$(W, 3)$; Append $\widetilde{A}$ to A

end for

return A