# Explainable Deepfake Detection: A Survey of Methods and Trends

**Lifu Huang, Yuanbiao Gou***
Sichuan University
{lifuhuang37, gouyuanbiao}@gmail.com

## Abstract

The rapid advancement of generative AI has accelerated the spread of Deepfake technologies, posing increasing threats to media authenticity, public safety, and legal forensics. While detection algorithms have made significant strides in accuracy, many remain opaque black-box models, offering little insight into their decision-making processes. This lack of interpretability undermines trust and hinders their use in high-stakes, real-world scenarios. This survey offers a concise overview of current and emerging approaches to explainable Deepfake detection, with a focus on enhancing transparency, verifiability, and human alignment. We categorize these methods into four main paradigms: visual localization techniques, feature attribution methods, natural language explanations, and agent-based interactive reasoning. In addition, we highlight key open challenges and propose future research directions, including weakly supervised explanation generation, unified multimodal reasoning, and real-world deployment. Our aim is to support the development of transparent, robust, and accountable Deepfake detection systems suitable for complex real-world environments.

## 1 Introduction

The rapid advancement of generative AI has led to an unprecedented surge in Deepfake content—synthetic media that convincingly mimics real humans and scenarios. Although early Deepfakes focused on facial manipulations [Rossler *et al.*, 2019], recent developments extend to multimodal synthetic content, including audio [Frank and Schönherr, 2021], text [Zellers *et al.*, 2019], and cross-modal generation [Chung *et al.*, 2017]. This evolution poses a significant threat to public trust, legal evidence chains, and digital integrity in critical scenarios.

Numerous detection models have been developed to address the threat of deepfakes; however, many of these function as black-box classifiers, offering only binary or probabilistic outputs without insight into their decision-making processes. Afchar et al. [2018] introduced MesoNet, a compact convolutional neural network designed for Deepfake detection. Although MesoNet achieves commendable performance on benchmark datasets, it lacks mechanisms to provide interpretable insights into its decision-making process. Li and Lyu [2018] proposed DSP-FWA, which detects Deepfake videos by identifying face warping artifacts using a dual spatial pyramid strategy. While effective in detection, DSP-FWA does not offer explanations for its classifications. Nguyen et al. [2019] presented a multi-task learning approach that simultaneously performs detection and localization of manipulated facial images. Despite its ability to highlight manipulated regions, the model does not provide comprehensive rationales for its decisions. Such outputs, though effective in benchmarking, often fall short in high-stakes real-world applications such as law enforcement, forensic auditing, and online content moderation. These domains demand more than a yes/no answer, they require interpretable, verifiable, and traceable evidence that can support human decision-making.

In response, a growing body of research has focused on enhancing the transparency of deepfake detection methods, aiming to bridge the gap between high-performing classifiers and systems that are interpretable and trustworthy to human users. Existing techniques include visual localization methods such as Grad-CAM [Selvaraju *et al.*, 2017] and saliency heatmaps [Simonyan *et al.*, 2013], which highlight discriminative regions to aid in interpreting model predictions. Recent works have also explored natural language explanations using large language models, as demonstrated by Jia et al. [Jia *et al.*, 2024]. Furthermore, multimodal reasoning that combine vision and language modalities, such as [Chakraborty *et al.*, 2025] and [Kundu *et al.*, 2025], have been proposed to generate coherent, interpretable evidence chains. Despite these promising developments, the field remains fragmented, with approaches emerging across diverse paradigms but lacking a cohesive taxonomy or shared conceptual foundation. A clearer organization of methods and deeper understanding of evolving trends are needed to guide future research toward interpretable and trustworthy Deepfake detection.

This paper presents a concise survey of recent advances in explainable deepfake detection. We systematically categorize both established and emerging approaches, and highlight key research challenges and gaps. By offering a structured and integrative perspective, this survey aims to support the de-

*Corresponding author

velopment of detection systems that are not only performant but also transparent, interpretable, and accountable, meeting the dual demands of algorithmic efficacy and human-centered trust. The main contributions of this work are as follows:

- We offer a structured taxonomy of explainability methods for Deepfake detection, categorizing the field into four functional paradigms, *i.e.*, visual localization, feature attribution, natural language explanation, and agent-based interactive reasoning.
- We highlight key research challenges and open questions, such as unified explainability frameworks, weakly supervised explanation, agent-based interactive systems, and real-world deployment, with the aim of guiding future work in building transparent and accountable detection systems.

## 2 Explainability Techniques in Deepfake Detection

Explainability in Deepfake detection refers to a model's ability to provide human-understandable reasons for its predictions, ideally in a form that not only indicates whether content is fake, but also how and where manipulations occur. Existing techniques can be broadly categorized into four major paradigms: visual explanation methods, feature attribution techniques, natural language explanations, and agent-based reasoning frameworks.

### 2.1 Visual Explanation and Localization Maps

One of the most intuitive and widely used approaches to explain Deepfake detection models is through spatial localization, highlighting manipulated regions in the input image or video that significantly influenced the model's prediction. These methods aim to provide visual justifications that are particularly useful in high-stakes applications.

A common class of techniques involves heatmap generation, such as Grad-CAM [Selvaraju *et al.*, 2017] and saliency maps [Simonyan *et al.*, 2013]. These methods compute class-specific gradient activations or pixel-wise relevance scores over convolutional layers, revealing spatial regions that contribute most to the model's final decision. In the context of Deepfake detection, they have been used to expose pixel-level artifacts introduced by face swapping, frame interpolation, or expression warping. For example, Silva et al. [2022] proposed a hierarchical ensemble network and used Grad-CAM visualizations to interpret the spatial regions associated with forged facial areas, enhancing model transparency for forensic analysis. Similarly, Dong et al. [2022] analyzed matching-based features extracted by a CNN-based Deepfake detector, and employed heatmap-based visualization to interpret how the model focuses on manipulated regions. Naskar et al. [2024] employed Grad-CAM to visualize the attention regions of their base CNN models, offering insights into how each model responds to manipulated facial areas. These visualizations support the interpretability of feature extraction and stacking in their ensemble framework.

Another line of work leverages attention-based visualization. Transformer-based models, such as Vision Transformers (ViT) [Dosovitskiy *et al.*, 2020], naturally encode spatial attention across image patches. By extracting attention weights or joint attention matrices, researchers can identify discriminative regions that the model attends to when detecting manipulations. Building on this idea, Nguyen et al. [2024] investigate the use of self-supervised ViTs for deepfake detection. They show that partial fine-tuning of the final transformer blocks allows the model to adapt its attention specifically to manipulation cues. Notably, their experiments demonstrate that the fine-tuned DINOv2 models consistently attend to semantically meaningful facial regions (e.g., eyes, nose, mouth) where deepfake artifacts are likely to appear. This not only improves detection accuracy but also enhances the explainability of the model, making attention maps a valuable interpretive tool in deepfake forensics. Extending this focus on explainability to other modalities, similar efforts are being made in audio deepfake detection. For instance, recent work introduces novel explainability methods for state-of-the-art transformer-based audio deepfake detectors [Channing *et al.*, 2024]. This research employs techniques such as attention roll-out to visualize and highlight the regions of the audio input that influence the model's decisions, thereby narrowing the explainability gap for human experts in the audio domain. Complementing these efforts, a practical detector is available as a GitHub repository [2022]. This detector, consist a Video Deepfake detector based on hybrid EfficientNet CNN and Vision Transformer architecture. The model results can be analyzed and explained by rendering a visualization based on a Relevancy map calculated from the Attention layers of the Transformer, overlayed on the input face image.

In addition, perturbation-based methods, such as occlusion sensitivity analysis [Yosinski *et al.*, 2015] or mask-based testing [Fong and Vedaldi, 2017], systematically modify input regions to observe changes in model outputs. These methods highlight causal image areas whose presence or absence strongly affects the classification, thereby offering empirical insight into the model's decision behavior. This line of work is also actively explored in the context of Deepfake detection. For instance, some research [Tsigos *et al.*, 2024] specifically focuses on evaluating the effectiveness of various Explainable AI methods, including perturbation-based approaches, for Deepfake detection models. Furthermore, similar techniques are applied to other modalities. For audio deepfake detection, explainability methods that employ occlusion have been introduced to highlight the specific regions of the audio input influencing the model's decisions, thus aiming to close the explainability gap for detectors [Channing *et al.*, 2024].

Overall, these visual explanation methods are essential for producing granular, region-level evidence that goes beyond a mere binary classification of real or fake. By visually pinpointing the specific features or artifacts that a Deepfake detection model identifies as indicative of manipulation, these methods enable human auditors and legal experts to verify or contest algorithmic decisions in a transparent and interpretable manner. This transparency is crucial for building trust in automated systems, especially in high-stakes applications like media authentication and forensic analysis. It allows human oversight, facilitates accountability for the detection results, and provides actionable insights that can be vital

Table 1: Explainability Techniques for Deepfake Detection

| Paradigm | Key Techniques | Descriptions | Capabilities |
|---|---|---|---|
| Visual Explanation and Localization Maps | Grad-CAM, Saliency Maps, Attention Maps, Occlusion Sensitivity | Highlights spatial regions in images, videos, or spectrograms that contribute to the model's decision, revealing where manipulations likely occur. | Visualizes forged regions; supports image and audio spectrograms |
| Feature Attribution and Token-Level Insights | DeepLIFT, Integrated Gradients, SHAP, LRP, SOBOL | Quantifies the contribution of specific input features (pixels, patches, or tokens) to predictions, enabling fine-grained, interpretable analysis. | Provides pixel-, patch-, and token-level attribution |
| Natural Language Explanation with LLMs | Prompt-based Fine-tuning of LLMs, Vision-language Co-training | Generates textual explanations that describe why content is considered fake, which parts are suspicious, and what cues the model used to decide. | Produces contextual, user-friendly explanations |
| Agent-based Interactive Explanation | User-Agent Interaction, Debate Agents (Agent-Agent Interaction) | Uses agents capable of reasoning and engaging in multi-turn dialogues with agents/users to explain or justify detection outcomes. | Emulates human reasoning; supports dynamic user queries |

for understanding the nature of the deepfake and its potential impact.

## 2.2 Feature Attribution and Token-Level Insights

Beyond visual localization, feature attribution represents a complementary paradigm for explaining deepfake detection models. These methods aim to quantify the individual contributions of specific input features or tokens to a model's prediction confidence. By assigning importance scores to discrete data fragments or their abstract representations, feature attribution techniques offer a more fine-grained understanding of model behavior.

One prominent method in this category is Integrated Gradients (IG)[Sundararajan *et al.*, 2017]. As a path-based attribution technique, IG computes the integral of gradients along a linear path from a baseline input to the actual input, thereby quantifying each feature's contribution to the model's prediction. In the context of deepfake detection, IG enables attribution of decisions to specific input features, such as pixels in images or time-frequency patterns in audio, offering a theoretically grounded interpretation of model behavior. For example, IG has been employed in deepfake voice detection to analyze model predictions with respect to perceptually meaningful features, such as formant structures in spectrograms [Lim *et al.*, 2022]. Beyond unimodal settings, IG has also been extended to multimodal detection scenarios. In the Multimodaltrace framework [Raza and Malik, 2023], IG is applied to jointly interpret audio and visual inputs, providing insight into how the model differentially attends to modalities when making predictions across multiple classification heads. Such applications demonstrate IG's versatility and its potential for delivering human-understandable explanations in complex, multimodal detection systems.

SHAP [Lundberg and Lee, 2017] is a model-agnostic attribution method that offers a semantic interpretation of model behavior by quantifying the contribution of each input feature to the final prediction. Rooted in cooperative game theory, SHAP conceptualizes input features, such as pixels, audio tokens, or facial landmarks, as players in a coalition game. The model's prediction serves as the "payout," which is fairly distributed among features using Shapley values. These values are computed by systematically perturbing combinations of inputs and observing changes in model output, thereby isolating each feature's marginal contribution. In deepfake detection, SHAP has been used to interpret both traditional classifiers and deep neural networks, highlighting which input elements (e.g., facial regions, texture inconsistencies, or spectral artifacts) are most influential in classification decisions. For instance, SHAP has been applied to spoofing and speech-based deepfake detection [Ge *et al.*, 2022], where it revealed non-obvious model behaviors and localized critical artifacts. Moreover, SHAP is often included in comparative evaluations of explainable AI techniques for deepfake detection [Tsigos *et al.*, 2024], particularly those involving black-box models trained on benchmarks like FaceForensics++. These studies underscore SHAP's utility in providing faithful and human-interpretable explanations across diverse detection scenarios.

For models that process inputs as discrete units, such as Vision Transformers (ViTs), which divide images into fixed-size patches, patch-level attribution has emerged as a particularly relevant technique. In these architectures, the input image is partitioned into a grid of patches, each treated as a token within the attention mechanism. Attribution methods such as DeepLIFT [Shrikumar *et al.*, 2017] and LRP [Bach *et al.*, 2015] can be applied at the patch level, enabling the assignment of importance scores to individual patches. This fine-grained analysis facilitates the localization of suspicious regions likely to contain deepfake artifacts, for instance, the mouth area in lip-sync manipulations or the eyes in expression-forged videos. Nguyen et al. [2024] demonstrate that fine-tuned Vision Transformers tend to focus their attention on semantically meaningful facial regions where manipulations are most likely to occur. By examining attention maps or applying attribution techniques at the patch level, it

becomes possible to precisely identify and visualize these regions, offering interpretable insights into the model's detection rationale.

Another notable attribution-based method is SOBOL [Fel *et al.*, 2021], which leverages Sobol' indices from sensitivity analysis to quantify the contribution of input variables to the variance in model outputs. The technique operates by generating real-valued perturbation masks, sampled using Quasi-Monte Carlo sequences, and applying them to the input image through functions such as blurring. By measuring how these perturbed inputs influence the model's predictions, SOBOL estimates total-order Sobol' indices, producing visual explanations that highlight the most influential input regions. In a recent study, Tsigos et al. [2024] evaluated SOBOL alongside other explanation methods within a quantitative evaluation framework tailored to deepfake detection. Their findings indicated that SOBOL consistently ranked among the top-performing methods in localizing manipulated regions. Expanding on this work, Tsigos et al. [2025] introduced SOBOLadv, an enhanced variant that incorporates adversarially generated samples to create more realistic and semantically meaningful perturbation masks. This improved version demonstrated increased explanation accuracy and sufficiency, underscoring SOBOL's robustness and adaptability in the deepfake detection domain.

Beyond spatial features, deepfake detection, particularly in multimodal and sequential domains, also benefits from token-level attribution applied to abstract representations. In audio deepfake detection, raw signals are commonly transformed into sequences of mel-spectrograms or self-supervised embeddings. Attribution techniques assign importance scores to specific time-frequency bins or embedding tokens, illuminating which acoustic features, such as pitch variations, timbre shifts, or atypical speech patterns, contribute most to the detection of synthetic speech. Channing et al. [2024] tackle this challenge by proposing novel explainability methods tailored for transformer-based audio deepfake detectors. Their approach leverages spectrogram inputs to pinpoint time- and feature-specific regions critical for model decisions, facilitating a more interpretable and granular analysis of deepfake artifacts in the audio domain.

Overall, feature attribution and token-level insight techniques offer a granular and quantitative understanding of model decisions in deepfake detection. These methods go beyond merely highlighting where manipulations may occur, to elucidating which features drive the detection and to what extent they influence the outcome. Such detailed explanations are essential for multiple reasons: they help reveal potential model biases by identifying over-relied-upon regions or features; they inform model refinement by pinpointing areas requiring more robust representation learning; and they deepen human comprehension of the interplay between deepfake generation artifacts and detection mechanisms. This enhanced interpretability is critical for advancing more resilient, transparent, and trustworthy deepfake detection systems.

### 2.3 Natural Language Explanation with LLMs

Inspired by recent advances in Natural Language Processing (NLP) and Visual Question Answering (VQA), a growing trend in deepfake detection is the generation of textual justifications to accompany model predictions. Rather than restricting outputs to binary labels, this approach utilizes large language models (LLMs) and vision-language models (VLMs) to generate human-readable explanations that answer questions such as, "Why is this video considered fake?" or "Which parts appear unnatural, and why?".

The core methodologies in this emerging area combine the predictive capabilities of deepfake detectors with the generative strengths of LLMs. A common strategy is prompt-based finetuning, wherein multimodal embeddings, such as visual features from manipulated regions or audio cues from synthetic speech, are extracted by a detector and provided to an LLM alongside carefully crafted prompts. The LLM is then finetuned to produce coherent, context-aware natural language explanations that interpret the detector's outputs. This process effectively translates complex, high-dimensional feature representations into human-readable justifications that highlight the presence of anomalies. Guo et al. [2025] introduce a Multi-Modal Face Forgery Detector that performs both binary classification and explanation generation. Their approach integrates the multimodal representation learning of a pre-trained CLIP model with the interpretability of LLMs via customized prompt learning for face forgery. This enables the LLM to generate detailed textual explanations that align natural language descriptions with subtle indicators of visual manipulation. Similarly, Yu et al. [2025] explore the use of VLMs for generalizable and explainable deepfake detection. Their method constructs forgery-specific prompt embeddings, which are passed into an LLM trained to generate rich, context-sensitive explanations.

Another promising direction involves co-training vision-language architectures, where the Deepfake detection task is integrated directly into a vision-language model's training process. The primary goal is to align visual or auditory evidence of manipulation with corresponding textual explanations. This approach allows for training the VLM to simultaneously identify deepfakes and generate captions or descriptive text that precisely points out manipulated regions or artifacts. An example is TruthLens [Kundu *et al.*, 2025], which provides explainable deepfake detection by simultaneously classifying images as real or fake and generating detailed textual reasoning. This is achieved through a synergistic training process that combines the global contextual understanding of multi-modal large language models (such as PaliGemma2) with the localized feature extraction of vision-only models (like DINOv2). This integrated training aligns visual cues directly with textual explanations, enabling the framework to effectively handle both face-manipulated deepfakes and fully AI-generated content, and to address fine-grained queries (e.g., "Does the eyes/nose/mouth look real or fake?"). Another example is the work by Zhang et al. [2024], which re-frames deepfake detection as a Deepfake Detection Visual Question Answering (DD-VQA) task. The proposed Vision and Language Transformer-based framework is co-trained to provide both classification and textual explanations grounded in common sense reasoning. This is achieved by incorporating text- and image-aware feature alignment formulation to enhance multi-modal representation learning, di-

rectly linking visual evidence of forgery with descriptive textual justifications.

While still in an exploratory phase, this research direction holds significant promise for bridging the gap between complex deep learning models and human interpretability. In contrast to conventional tools such as heatmaps or attribution scores, natural language explanations provide a more intuitive and accessible way to convey model reasoning. They are capable of articulating nuanced decision logic, offering contextual insights, and even suggesting actionable interpretations—features that are especially valuable in forensic settings and public communication. Such interpretability is crucial for building trust in automated systems, particularly in the high-stakes domain of AI-generated content detection, where transparency and accountability are paramount. Looking forward, future research will likely focus on improving the factual reliability, clarity, and robustness of generated explanations, with an emphasis on minimizing hallucinations and adapting outputs for practical, real-world deployment.

### 2.4 Agent-based Interactive Explanation

Agent-based interactive explanation offers a promising pathway toward improving the explainability and transparency of deepfake detection. By leveraging modular, role-specific agents alongside dynamic agent–agent and user–agent interactions, these systems can more closely emulate human-like reasoning and forensic workflows, making them especially valuable in high-stakes domains such as journalism, legal analysis, and content moderation.

**Modular Pipelines.** A key advantage of agent-based systems lies in their ability to decompose the detection process into specialized, manageable subtasks. Rather than relying on a monolithic model, a multi-agent architecture delegates distinct roles to dedicated agents. For instance, a detection agent may first conduct authenticity classification to flag potentially manipulated media. Once a sample is identified as suspicious, a localization agent can analyze spatial or temporal inconsistencies, such as unnatural facial expressions or mismatched audio segments, thereby narrowing down the regions of interest. Finally, an explanation agent aggregates the findings and generates interpretable outputs in the form of heatmaps, textual justifications, or interactive forensic reports. This modular structure not only enhances analytical robustness and flexibility but also improves traceability and user trust, as the decision-making processes of individual agents can be independently examined and understood.

**Debate Agents.** Recent advances have shown that LLMs can be effectively integrated into multi-agent systems for deepfake detection. A notable example is the system proposed by Jeptoo and Sun [2024], which adopts a structured debate format involving LLM agents assigned distinct roles, such as fact-checkers, journalists, and data analysts. Through deliberative dialogue, these agents collaboratively assess content authenticity while simultaneously articulating the rationale behind their judgments. In a similar vein, Liu et al. [2025] introduced the TruEDebate (TED) framework, which incorporates DebateFlow and InsightFlow agents to present and challenge claims, uncover fabricated content, and support transparent decision-making. These debate agents provide an interpretable interface by exposing the reasoning chains, diverse perspectives, and counterarguments that underpin the final conclusions.

**User–Agent Interaction.** Although inter-agent interaction enhances the transparency of internal reasoning processes, user–agent/model interaction plays an equally critical role in improving system usability and fostering user trust. Traditional deepfake detectors often yield static outputs, limiting users' ability to explore the system's underlying rationale. In contrast, interactive systems support multi-turn dialogues, allowing users to ask follow-up questions, request clarifications, or investigate alternative explanations. Yu et al. [2025] introduce a VLM-based deepfake detection framework that integrates visual prompt embeddings, forgery-aware textual features, and natural language querying. Their system enables multi-turn interactions with users, facilitating interactive reasoning and localized explanations, thereby significantly improving transparency and user comprehension in deepfake forensics.

Despite their promise, agent-based systems also pose several challenges. Effective coordination among agents is crucial to avoid inefficiencies and bottlenecks, particularly as data flows through multiple analytical stages. Latency may become an issue, especially in real-time or high-volume settings. Furthermore, error propagation represents a significant risk: a misclassification by an upstream agent, such as incorrectly flagging an authentic video as fake, can distort downstream analyses and ultimately reduce the interpretability and reliability of the overall system.

In conclusion, agent-based interactive explanation represents a compelling paradigm for advancing explainable deepfake detection. By adopting a modular architecture that decomposes the detection process into specialized, interacting components, these frameworks enable more transparent, scalable, and accountable forensic workflows. Looking ahead, future research should prioritize improving inter-agent coordination, enhancing user interactivity, and strengthening the robustness and fidelity of explanations to fully unlock the potential of these architectures in real-world deployments.

## 3 Open Challenges and Future Directions

Despite encouraging progress in explainable deepfake detection, several critical research challenges remain. This section outlines key obstacles and promising avenues for advancing the field toward more reliable, scalable, and human-aligned interpretability.

### 3.1 Unified Explainability Frameworks for Multimodal Deepfakes

While most existing explainability methods have been developed for single-modal deepfakes, recent advances in generative models have enabled the creation of multimodal deepfakes that blend video, audio, text, and even behavioral cues. This evolution introduces new challenges: current explainability techniques often operate in isolation, failing to capture the intricate interplay of manipulation signals across modalities. As a result, their utility is limited in real-world scenarios, where multimodal forgeries are increasingly prevalent.

To address this gap, future research must move toward modality-agnostic explanation frameworks that deliver holistic and interpretable assessments of deepfake content. A unified approach should address three key components: i) Modality-specific attribution: First, the system must detect and attribute manipulation cues unique to each modality. For example, unnatural prosody in audio, lip-sync inconsistencies in video, or contradictions between visual and textual content. ii) Cross-modal reasoning: These diverse signals must then be integrated through advanced reasoning mechanisms capable of evaluating their interdependence and synthesizing a coherent explanation of the content's authenticity. iii) Multiform explanation output: To ensure accessibility and usability, explanations should be presented in both visual and textual formats. For instance, combining attention heatmaps that localize artifacts with narrative descriptions that contextualize their significance.

Vision-language models (VLMs) and cross-modal attention mechanisms offer promising foundations for building such systems. VLMs can align visual evidence with linguistic reasoning, enabling rich, human-readable justifications. Meanwhile, cross-modal attention techniques can help prioritize and unify signals across modalities, producing more focused and accurate interpretations of complex manipulations. Together, these technologies may support the next generation of explainable systems that are robust, multimodal, and aligned with human interpretive needs.

### 3.2 Weakly- and Unsupervised Explanation

Many real-world deepfake detection scenarios lack access to dense supervision, such as pixel-level manipulation masks or human-authored explanation annotations. Relying on fully annotated datasets constrains the scalability and adaptability of explainable detection systems, as manual labeling is labor-intensive, time-consuming, and often error-prone. To address these limitations, future research must explore weakly supervised and unsupervised approaches for explanation generation, enabling broader deployment of interpretable systems in diverse, data-scarce environments.

Learning manipulation patterns via anomaly detection or self-supervised objectives. Instead of depending on explicit labels, models can learn to identify deepfake characteristics by detecting deviations from the statistical patterns of authentic content. Self-supervised learning can leverage intrinsic structures in the data, such as temporal coherence or frequency regularity, to learn meaningful representations that implicitly encode manipulation cues.

Aligning visual signals with textual outputs using minimal human feedback. Techniques that can infer relationships between visual anomalies and their textual descriptions with limited even no direct human input are crucial. This could involve using contrastive learning or alignment methods to bridge the gap between visual and linguistic representations.

Employing prompt engineering with LLMs for zero-shot or few-shot explanation generation. The emergent reasoning capabilities of LLMs can be harnessed through carefully crafted prompts to synthesize coherent textual explanations from minimal visual cues. This approach avoids task-specific finetuning, enabling flexible and low-cost generation of explanatory outputs.

Adopting weakly- and unsupervised techniques would significantly lower the dependency on labeled data, improve generalization across modalities and manipulation types, and facilitate scalable deployment of explainable deepfake detection systems in real-world applications where annotated resources are limited.

### 3.3 Agent-based Interactive Explanation

A promising yet underexplored direction in deepfake detection is the development of modular, agent-based interactive systems. In contrast to traditional monolithic models that often function as black boxes, modular architectures decompose the detection pipeline into discrete, cooperative agents, each assigned a specialized subtask such as detection, localization, or explanation. This structural design introduces several key advantages for interpretability and adaptability.

Inherent interpretability by design. Each agent is responsible for a clearly defined analytical function, such as detecting facial inconsistencies, identifying audio-visual mismatches, or aligning multimodal evidence. This division of labor facilitates fine-grained traceability, allowing system outputs to be linked back to specific reasoning components.

Step-by-step, traceable rationales. Modular pipelines naturally produce a sequence of intermediate outputs, from initial feature extraction to final classification. These outputs can be aggregated into layered, transparent ("white-box") explanations, supporting greater user understanding and diagnostic insight into the decision process.

Scalability across manipulation types. As deepfake generation techniques continue to diversify, modular frameworks offer greater flexibility. New agents can be introduced or existing ones fine-tuned to handle emerging modalities or manipulation strategies, without retraining the entire system, enabling continuous system evolution.

To further enhance reasoning capabilities and interactivity, large foundation models such as LLMs and VLMs can serve as high-level controller modules. These models can orchestrate agent behavior, unify heterogeneous signals, and generate coherent natural language justifications. This hybrid approach, combining modular specialization with the generalization power of foundation models, presents a scalable, extensible, and human-aligned pathway for interactive and explainable deepfake detection.

### 3.4 Real-world Deployment and Societal Alignment

Ultimately, explainable Deepfake detectors must function effectively in real-world settings, where inputs are noisy, users are non-experts, and the implications of decisions can be far-reaching. The transition from controlled experimental environments to real-world deployment introduces a host of challenges that extend well beyond algorithmic accuracy, encompassing ethical, legal, and societal considerations. Several pressing questions and emerging risks must be addressed:

How do the explanations influence human trust, decision-making, and accountability? While explanations are intended to build trust, poorly designed or misleading explanations

could have the opposite effect. It is crucial to understand how different types of explanations influence user confidence, whether they lead to more accurate human judgments, and how accountability for decisions made with the aid of these systems is distributed.

Can the explanations themselves be adversarially manipulated? Recent studies in explainable AI have shown that explanation outputs can be targeted and altered without affecting model predictions, a phenomenon known as explanation manipulation. In the Deepfake context, a malicious actor might craft adversarial examples that not only bypass detection but also generate plausible-looking heatmaps or rationales, deliberately misleading the user.

How do explanations propagate or mask algorithmic bias? Explanations are not inherently neutral, they often reflect and can reinforce biases present in a model's training data or design. A deepfake detection system trained largely on Western celebrity datasets may underperform, and produce misleading explanations, when analyzing individuals from underrepresented ethnic or cultural groups. This raises equity concerns that biased explanations may wrongly flag content from marginalized communities or fail to justify the authenticity of real content, reinforcing mistrust and exclusion.

How to balance transparency with privacy and security concerns? While interpretability often requires surfacing internal model processes, doing so may inadvertently reveal proprietary methods or system vulnerabilities. Attackers could exploit explanation outputs to reverse-engineer detection logic, thus crafting more evasive Deepfakes. Moreover, certain explanations might expose sensitive metadata or individual biometric features, creating privacy risks. A careful balance must be struck between offering sufficient transparency and safeguarding privacy and security.

These concerns underscore the need for ethical, legal, and policy frameworks developed in parallel with technological solutions. Responsible deployment of explainable deepfake detection demands interdisciplinary collaboration, engaging computer scientists, sociologists, legal experts, ethicists, and affected communities. Priorities include defining metrics for explanation robustness and fairness, creating oversight mechanisms for disputed outcomes, and adopting user-centered design standards that tailor explanations to diverse audiences. Without such safeguards, even technically sound systems risk eroding public trust and amplifying harm.

## 4 Conclusion

As Deepfake technologies grow in realism, scale, and modality, the demand for explainable and trustworthy detection systems becomes increasingly urgent. This survey reviewed recent advances across four key paradigms: visual localization, feature attribution, natural language explanation, and agent-based interactive reasoning. Each contributes a distinct lens on interpretability, *i.e.*, from spatial and pixel-level evidence to human-readable justifications with multi-agent collaboration.

Despite notable progress, major challenges persist, including limited cross-modal integration, annotation scarcity, and deployment in real-world settings. We outlined several future directions to address these gaps, involoving unified multimodal frameworks, weakly supervised explanation methods, modular agent-based architectures, and ethically grounded deployment strategies.

Ultimately, explainability is not only a technical enhancement but a foundation for accountability, transparency, and public trust in high-stakes domains. Realizing this vision requires the development of scalable, human-aligned models, and sustained interdisciplinary collaboration among technologists, domain experts, and affected communities.

## References


[Afchar *et al.*, 2018] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[Bach *et al.*, 2015] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[Chakraborty *et al.*, 2025] Ritabrata Chakraborty, Rajatsubhra Chakraborty, Ali Khaleghi Rahimian, and Thomas MacDougall. Truthlens: A training-free paradigm for deepfake detection. *arXiv preprint arXiv:2503.15342*, 2025.

[Channing *et al.*, 2024] Georgia Channing, Juil Sock, Ronald Clark, Philip Torr, and Christian Schroeder de Witt. Toward robust real-world audio deepfake detection: Closing the explainability gap. *arXiv preprint arXiv:2410.07436*, 2024.

[Chung *et al.*, 2017] Joon Son Chung, Amir Jamaludin, and Andrew Zisserman. You said that? *arXiv preprint arXiv:1705.02966*, 2017.

[Dong *et al.*, 2022] Shichao Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. Explaining deepfake detection by analysing image matching. In *European conference on computer vision*, pages 18–35. Springer, 2022.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fel *et al.*, 2021] Thomas Fel, Rémi Cadène, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. *Advances in neural information processing systems*, 34:26005–26014, 2021.

[Fong and Vedaldi, 2017] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017.

[Frank and Schönherr, 2021] Joel Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint arXiv:2111.02813*, 2021.

[Ge *et al.*, 2022] Wanying Ge, Jose Patino, Massimiliano Todisco, and Nicholas Evans. Explaining deep learning models for spoofing and deepfake detection with shapley additive explanations. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 6387–6391. IEEE, 2022.

[Guo *et al.*, 2025] Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. *arXiv preprint arXiv:2503.20188*, 2025.

[Jeptoo and Sun, 2024] Korir Nancy Jeptoo and Chengjie Sun. Enhancing fake news detection with large language models through multi-agent debates. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 474–486. Springer, 2024.

[Jia *et al.*, 2024] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4324–4333, 2024.

[Kundu *et al.*, 2025] Rohit Kundu, Athula Balachandran, and Amit K Roy-Chowdhury. Truthlens: Explainable deepfake detection for face manipulated and fully synthetic data. *arXiv preprint arXiv:2503.15867*, 2025.

[Li and Lyu, 2018] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

[Lim *et al.*, 2022] Suk-Young Lim, Dong-Kyu Chae, and Sang-Chul Lee. Detecting deepfake voice using explainable deep learning techniques. *Applied Sciences*, 12(8):3926, 2022.

[Liu *et al.*, 2025] Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. *arXiv preprint arXiv:2505.08532*, 2025.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[Naskar *et al.*, 2024] Gourab Naskar, Sk Mohiuddin, Samir Malakar, Erik Cuevas, and Ram Sarkar. Deepfake detection using deep feature stacking and meta-learning. *Heliyon*, 10(4), 2024.

[Nguyen *et al.*, 2019] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–8. IEEE, 2019.

[Nguyen *et al.*, 2024] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Exploring self-supervised vision transformers for deepfake detection: A comparative analysis. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2024.

[noame12, 2022] noame12. Explainable attention based deepfake detector. https://github.com/noame12/Explainable_Attention_Based_Deepfake_Detector, 2022.

[Raza and Malik, 2023] Muhammad Anas Raza and Khalid Mahmood Malik. Multimodaltrace: Deepfake detection using audiovisual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 993–1000, 2023.

[Rossler *et al.*, 2019] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Shrikumar *et al.*, 2017] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR, 2017.

[Silva *et al.*, 2022] Samuel Henrique Silva, Mazal Bethany, Alexis Megan Votto, Ian Henry Scarff, Nicole Beebe, and Peyman Najafirad. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Science International: Synergy*, 4:100217, 2022.

[Simonyan *et al.*, 2013] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[Tsigos *et al.*, 2024] Konstantinos Tsigos, Evlampios Apostolidis, Spyridon Baxevanakis, Symeon Papadopoulos, and Vasileios Mezaris. Towards quantitative evaluation of explainable ai methods for deepfake detection. In *Proceedings of the 3rd ACM international workshop on multimedia AI against disinformation*, pages 37–45, 2024.

[Tsigos *et al.*, 2025] Konstantinos Tsigos, Evlampios Apostolidis, and Vasileios Mezaris. Improving the perturbation-based explanation of deepfake detectors through the use of adversarially-generated samples. *arXiv preprint arXiv:2502.03957*, 2025.

[Yosinski *et al.*, 2015] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[Yu *et al.*, 2025] Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip Hong Chang. Unlocking the capabilities of vision-language models for generalizable and explainable deepfake detection. *arXiv preprint arXiv:2503.14853*, 2025.

[Zellers *et al.*, 2019] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *Advances in neural information processing systems*, volume 32, 2019.

[Zhang *et al.*, 2024] Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *European Conference on Computer Vision*, pages 399–415. Springer, 2024.