Novelty Assessment Report

Paper: FakeXplain: AI-Generated Images Detection via Human-Aligned Grounded Reasoning
PDF URL: https://openreview.net/pdf?id=UcpTOa8OnG
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

The rapid rise of image generation calls for detection methods that are both interpretable and reliable. Existing approaches, though accurate, act as black boxes and fail to generalize to out-of-distribution data, while multi-modal large language models (MLLMs) provide reasoning ability but often hallucinate. To address these issues, we construct FakeXplained dataset of AI-generated images annotated with bounding boxes and descriptive captions that highlight synthesis artifacts, forming the basis for human-aligned, visually grounded reasoning. Leveraging FakeXplained, we develop FakeXplainer which fine-tunes MLLMs with a progressive training pipeline, enabling accurate detection, artifact localization, and coherent textual explanations. Extensive experiments show that FakeXplainer not only sets a new state-of-the-art in detection and localization accuracy (98.2% accuracy, 36.0% IoU), but also demonstrates strong robustness and out-of-distribution generalization, uniquely delivering spatially grounded, human-aligned rationales.

Core Task Landscape

This paper addresses: **AI-Generated Image Detection with Interpretable Explanations**

A total of **50 papers** were analyzed and organized into a taxonomy with **20 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Multimodal Large Language Model-Based Detection and Explanation**
- **Deep Learning Classification Approaches**
- **Explainable AI Techniques for Detection Transparency**
- **Forensic and Signal-Based Detection Methods**
- **Datasets, Benchmarks, and Evaluation Frameworks**
- **User-Centric and Interactive Detection Systems**
- **Domain-Specific and Specialized Applications**
- **Generalization, Robustness, and Cross-Generator Detection**

Complete Taxonomy Tree

- AI-Generated Image Detection with Interpretable Explanations Survey Taxonomy
- Multimodal Large Language Model-Based Detection and Explanation
  - Grounded Reasoning with Artifact Localization ★ (5 papers)
  - [0] FakeXplain: AI-Generated Images Detection via Human-Aligned Grounded Reasoning (Anon et al., 2026) View paper
  - [3] ForenX: Towards Explainable AI-Generated Image Detection with Multimodal Large Language Models (Tan, 2025) View paper
  - [10] AIGI-Holmes: Towards Explainable and Generalizable AI-Generated Image Detection via Multimodal Large Language Models (Zhou Ziyin, 2025) View paper
  - [11] Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation (Wen Si-wei, 2025) View paper
  - [29] Interpretable and Reliable Detection of AI-Generated Images via Grounded Reasoning in MLLMs (Ji, 2025) View paper
  - Textual Reasoning Without Spatial Grounding (5 papers)
  - [4] Towards explainable fake image detection with multi-modal large language models (Ji, 2025) View paper
  - [17] Seeing before reasoning: A unified framework for generalizable and explainable fake image detection (Lin Kai-qing, 2025) View paper
  - [18] Fakebench: Probing explainable fake image detection via large multimodal models (Yixuan Li, 2025) View paper
  - [43] ThinkFake: Reasoning in Multimodal Large Language Models for AI-Generated Image Detection (Huang Tai-ming, 2025) View paper
  - [47] TruthLens:A Training-Free Paradigm for DeepFake Detection (Chakraborty Ritabrata, 2025) View paper
  - Cross-Modal and Multimodal Content Detection (4 papers)
  - [2] Sida: Social media image deepfake detection, localization and explanation with large multimodal model (Zhenglin Huang, 2025) View paper
  - [16] Decoding synthetic news: an interpretable multimodal framework for the classification of news articles in a novel news corpus (Michael Schlee, 2025) View paper
  - [22] BusterX++: Towards Unified Cross-Modal AI-Generated Content Detection and Explanation with MLLM (Wen Hai-quan, 2025) View paper
  - [45] METER: Multi-modal Evidence-based Thinking and Explainable Reasoning--Algorithm and Benchmark (Xu Yang, 2025) View paper
  - Training-Free and VQA-Based Paradigms (2 papers)
  - [49] X2-dfd: A framework for explainable and extendable deepfake detection (Chen Yize, 2024) View paper

    - [50] Using multimodal foundation models for detecting fake images on the internet with explanations (Vishnu S. Pendyala, 2024) View paper
- Deep Learning Classification Approaches
    - Convolutional Neural Network Architectures (6 papers)
    - [1] Detecting deepfake images using deep learning techniques and explainable AI methods (Wahidul Hasan Abir, 2023) View paper
    - [5] Detecting AI-generated images with CNN and Interpretation using Explainable AI (Bharathi Mohan G, 2024) View paper
    - [7] Cifake: Image classification and explainable identification of ai-generated synthetic images (Jordan J. Bird, 2024) View paper
    - [8] Detection of AI-generated synthetic images with a lightweight CNN (Adrian Lokner LaÄeviÄ, 2024) View paper
    - [33] Advancing AI-Generated Image Detection: Enhanced Accuracy through CNN and Vision Transformer Models with Explainable AI Insights (Md.Zahid Hossain, 2023) View paper
    - [35] Improving AI Generated Image Detection through an Interpretable and Enhanced CNN2D Architecture (Sujani, 2025) View paper
    - Vision Transformer-Based Detection (2 papers)
    - [15] Advanced detection of ai-generated images through vision transformers (Lamichhane, 2024) View paper
    - [21] Overcoming diagnostic and data privacy challenges in viral disease detection: an integrated approach using generative AI, vision transformers, explainable AI, and federated learning (Asadi Srinivasulu, 2025) View paper
    - Hybrid CNN-LSTM and Temporal Models (3 papers)
    - [13] Deepfake Image Detection Using Explainable AI and Deep Learning (Abdulrahman, 2025) View paper
    - [14] DeepExplain: enhancing deepfake detection through transparent and explainable AI model (A. Srinagesh, 2024) View paper
    - [32] Explainable AI for Deepfake Detection: A Grad-CAM Approach to Video Forensics (Jaya Lakshmi Narayana Budati, 2025) View paper
    - Lightweight and Efficient Architectures (1 papers)
    - [34] YOLOv8 framework for COVID-19 and pneumonia detection using synthetic image augmentation (Abdul Hasib Uddin, 2025) View paper
- Explainable AI Techniques for Detection Transparency
    - Gradient-Based and Saliency Visualization (3 papers)
    - [24] Extracting local information from global representations for interpretable deepfake detection (Elahe Soltandoost, 2025) View paper
    - [26] Explainable deep-fake detection using visual interpretability methods (Badhrinarayan Malolan, 2020) View paper
    - [36] Explainable AI for deepfake detection (Nazneen Mansoor, 2025) View paper
    - Perturbation-Based Explanation Methods (2 papers)
    - [19] Towards Quantitative Evaluation of Explainable AI Methods for Deepfake Detection (Konstantinos Tsigos, 2024) View paper
    - [37] Improving the perturbation-based explanation of deepfake detectors through the use of adversarially-generated samples (Apostolidis Evlampios, 2025) View paper
    - Network Dissection and Interpretable Architectures (1 papers)
    - [28] Enhancing Interpretability in AI-Generated Image Detection with Genetic Programming (Mingqian Lin, 2023) View paper
- Forensic and Signal-Based Detection Methods
    - Frequency and Spectral Domain Analysis (1 papers)
    - [9] MaskSim: Detection of synthetic images by masked spectrum similarity analysis (Yanhao Li, 2024) View paper
    - Semantic and Common-Sense Reasoning (1 papers)
    - [27] RADAR: Reasoning AI-Generated Image Detection for Semantic Fakes (Haochen Wang, 2025) View paper
- Datasets, Benchmarks, and Evaluation Frameworks
    - Large-Scale Annotated Detection Datasets (3 papers)
    - [23] DDL: A Large-Scale Datasets for Deepfake Detection and Localization in Diversified Real-World Scenarios (Miao, 2025) View paper
    - [30] DDL: A Dataset for Interpretable Deepfake Detection and Localization in Real-World Scenarios (Miao, 2025) View paper
    - [31] AiGen-FoodReview: A Multimodal Dataset of Machine-Generated Restaurant Reviews and Images on Social Media (Alessandro Gambetti, 2024) View paper
    - Benchmark and Evaluation Protocols (3 papers)
    - [39] The Visual Counter Turing Test (VCT2): A Benchmark for Evaluating AI-Generated Image Detection and the Visual AI Index (VAI) (Nasrin Imanpour, 2024) View paper
    - [40] AI-Generated Image Detection: An Empirical Study and Future Research Directions (Tasnim Nusrat, 2025) View paper
    - [42] Performance comparison and visualization of ai-generated-image detection methods (Daeeol Park, 2024) View paper
    - Synthetic Datasets for XAI Evaluation (1 papers)
    - [6] Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods (Robin Hesse, 2023) View paper
- User-Centric and Interactive Detection Systems (2 papers)
    - [12] LayLens: Improving Deepfake Understanding through Simplified Explanations (Abhijeet Narang, 2025) View paper
    - [41] From Prediction to Explanation: Multimodal, Explainable, and Interactive Deepfake Detection Framework for Non-Expert Users (Tariq Shahroz, 2025) View paper
- Domain-Specific and Specialized Applications
    - Sign Language and Specialized Visual Content (1 papers)
    - [20] Generation and Detection of Sign Language Deepfakes - A Linguistic and Visual Analysis (Shahzeb Naeem, 2024) View paper
    - Face Forgery and Identity-Aware Detection (2 papers)
    - [25] TruthLens: Explainable DeepFake Detection for Face Manipulated and Fully Synthetic Data (Kundu, 2025) View paper
    - [44] Identity-aware vision-language model for explainable face forgery detection (Xu Junhao, 2025) View paper
- Generalization, Robustness, and Cross-Generator Detection (3 papers)
    - [38] Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis (Reza Babaei, 2025) View paper
    - [46] Human-like Content Analysis for Generative AI with Language-Grounded Sparse Encoders (Tang YiMing, 2025) View paper
    - [48] The Art of Detection: Methods for Identifying AI-Generated Visual Content (Prof. Vaishali Suryawanshi, 2025) View paper

Narrative

Core task: AI-generated image detection with interpretable explanations. The field has evolved from purely classification-driven approaches toward systems that not only identify synthetic content but also provide human-understandable rationales for their decisions.

The taxonomy reflects this dual emphasis through eight major branches. Deep Learning Classification Approaches and Forensic and Signal-Based Detection Methods focus on discriminative accuracy using CNNs, vision transformers, and frequency-domain analysis. Explainable AI Techniques for Detection Transparency leverage saliency maps, attention mechanisms, and post-hoc interpretation tools to reveal which image regions or features drive predictions. Multimodal Large Language Model-Based Detection and Explanation represents a newer direction that integrates vision-language models to generate natural-language justifications and localize artifacts. Meanwhile, Datasets, Benchmarks, and Evaluation Frameworks establish standardized testbeds, User-Centric and Interactive Detection Systems explore human-in-the-loop workflows, Domain-Specific and Specialized Applications address niche contexts like medical imaging or sign language, and Generalization, Robustness, and Cross-Generator Detection tackle the challenge of maintaining performance across diverse generative models.

Recent work has increasingly emphasized grounded reasoning that pinpoints suspicious artifacts rather than offering only global verdicts. FakeXplain[0] exemplifies this trend by combining multimodal large language models with explicit artifact localization, situating itself within the Grounded Reasoning with Artifact Localization cluster. This approach contrasts with earlier explainability efforts such as Deepfake Detection Explainable[1] and CNN Explainable Detection[5], which primarily relied on gradient-based saliency or attention overlays without structured linguistic explanations. Closely related works like ForenX[3] and AIGI Holmes[10] similarly pursue fine-grained localization and interpretable outputs, yet FakeXplain[0] distinguishes itself by leveraging the reasoning capabilities of large language models to articulate why specific regions appear synthetic. A key open question across these branches is how to balance detection accuracy with explanation fidelity, especially when models must generalize to unseen generators or adversarially perturbed images.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. ForenX: Towards Explainable AI-Generated Image Detection with Multimodal Large Language Models

Authors: Tan, Chuangchuang, Wang Jing-lu, Chuangchuang Tan, Ming Xiang, et al. (18 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

Advances in generative models have led to AI-generated images visually indistinguishable from authentic ones. Despite numerous studies on detecting AI-generated images with classifiers, a gap persists between such methods and human cognitive forensic analysis. We present ForenX, a novel method that not only identifies the authenticity of images but also provides explanations that resonate with human thoughts. ForenX employs the powerful multimodal large language models (MLLMs) to analyze and int...

Relationship Analysis

Both papers belong to the same taxonomy category of grounded reasoning with artifact localization, using MLLMs to detect AI-generated images by identifying forgery artifacts and providing spatially grounded explanations. They overlap in their approach of fine-tuning MLLMs on human-annotated datasets with bounding boxes and textual descriptions of synthesis artifacts to enable interpretable detection. The key difference is that FakeXplain employs a progressive training pipeline with reinforcement learning (GRPO) and introduces a structured annotation framework with image-level tags, while ForenX focuses on forensic prompts to guide MLLM attention and introduces the ForgReason dataset curated through LLM-agent collaboration with human annotators.

---

2. AIGI-Holmes: Towards Explainable and Generalizable AI-Generated Image Detection via Multimodal Large Language Models

Authors: Zhou Ziyin, Luo YunPeng, Ziyin Zhou, Wu Yuan-chen, Yunpeng Luo, et al. (23 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

The rapid development of AI-generated content (AIGC) technology has led to the misuse of highly realistic AI-generated images (AIGI) in spreading misinformation, posing a threat to public information security. Although existing AIGI detection techniques are generally effective, they face two issues: 1) a lack of human-verifiable explanations, and 2) a lack of generalization in the latest generation technology. To address these issues, we introduce a large-scale and comprehensive dataset, Holmes-...

Relationship Analysis

Both papers belong to the Grounded Reasoning with Artifact Localization category, using MLLMs to detect AI-generated images by localizing forgery artifacts and providing textual explanations. They overlap in their approach of fine-tuning MLLMs on annotated datasets with bounding boxes and captions to achieve spatially grounded detection with human-aligned reasoning. The key difference is that FakeXplain focuses on a progressive training pipeline with GRPO reinforcement learning and emphasizes human-annotated bounding boxes, while AIGI-Holmes introduces a Multi-Expert Jury annotation method, a three-stage training framework with DPO, and a collaborative decoding strategy that integrates visual expert perception with MLLM semantic reasoning.

---

3. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation

Authors: Wen Si-wei, YE Junyan, Siwei Wen, Feng Peilin, Junyan Ye, et al. (20 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

With the rapid advancement of Artificial Intelligence Generated Content (AIGC) technologies, synthetic images have become increasingly prevalent in everyday life, posing new challenges for authenticity assessment and detection. Despite the effectiveness of existing methods in evaluating image authenticity and locating forgeries, these approaches often lack human interpretability and do not fully address the growing complexity of synthetic data. To tackle these challenges, we introduce FakeVLM, a...

Relationship Analysis

Both papers belong to the Grounded Reasoning with Artifact Localization category, using MLLMs to detect AI-generated images by localizing forgery artifacts and providing textual explanations. They overlap in their core approach of fine-tuning MLLMs on datasets with spatial annotations (bounding boxes) and natural language descriptions of synthesis artifacts to achieve interpretable detection. The key difference is that FakeXplain emphasizes human-aligned annotations through expert annotators and uses a progressive RL fine-tuning pipeline (SFT + GRPO) to achieve state-of-the-art grounding accuracy (36.0% IoU), while the candidate paper (FakeVLM/Spot the Fake) relies on multi-LMM automated annotation with category-specific knowledge injection and focuses on broader coverage across seven image categories including specialized types like satellite and document images.

---

4. Interpretable and Reliable Detection of AI-Generated Images via Grounded Reasoning in MLLMs

Authors: Ji, Yikun, Yan Hong, Yikun Ji, Lan Jun, et al. (18 authors total) | **Year/Venue**: 2025 • arXiv.org | **URL**: View paper

Abstract

The rapid advancement of image generation technologies intensifies the demand for interpretable and robust detection methods. Although existing approaches often attain high accuracy, they typically operate as black boxes without providing human-understandable

justifications. Multi-modal Large Language Models (MLLMs), while not originally intended for forgery detection, exhibit strong analytical and reasoning capabilities. When properly fine-tuned, they can effectively identify AI-generated image...

⚠ **Similarity Notice**

This paper appears to be a variant or near-duplicate of the original FakeXplain paper. Both describe the same FakeXplained dataset with 8,772 AI-generated images annotated with bounding boxes and captions, employ the same FakeXplainer model with progressive fine-tuning of MLLMs, and report identical or nearly identical performance metrics (98.2% accuracy, 36.0% IoU). The core technical contributions, methodology, and experimental setup are essentially identical, suggesting these are different versions of the same work.

Contributions Analysis

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: FakeXplained dataset with human-aligned grounded annotations

Description: A curated dataset of 8,772 AI-generated images from diverse state-of-the-art generative models, annotated with bounding boxes and concise captions that highlight visual anomalies and illogical details. This dataset provides fine-grained, human-grounded annotations to support both visual grounding and textual reasoning for interpretable detection.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. M3DSYNTH: A dataset of medical 3D images with AI-generated local manipulations

URL: View paper

Brief Assessment

M3DSYNTH[52] focuses on medical 3D CT images with manipulated lung nodules for diagnostic tampering detection, not general AI-generated image synthesis artifacts with visual grounding annotations for interpretable detection.

---

2. GeneVA: A Dataset of Human Annotations for Generative Text to Video Artifacts

URL: View paper

Brief Assessment

GeneVA[58] focuses on spatio-temporal artifacts in AI-generated videos with bounding box annotations, while the original paper addresses AI-generated images with grounded explanations. These are fundamentally different modalities and problem domains.

---

3. ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection

URL: View paper

Brief Assessment

ToxiGen[55] focuses on toxic language detection with machine-generated text about minority groups, not AI-generated image synthesis artifacts. The datasets serve entirely different domains (text vs. images) and tasks (hate speech detection vs. visual anomaly detection).

---

4. Efficient end-to-end learning for cell segmentation with machine generated weak annotations

URL: View paper

Brief Assessment

Cell Segmentation Learning[56] focuses on cell segmentation from microscopy images with machine-generated weak annotations (image-level segmentations and location-of-interests), not on AI-generated image detection with grounded visual anomaly annotations.

---

5. Cifake: Image classification and explainable identification of ai-generated synthetic images

URL: View paper

Brief Assessment

Cifake[7] focuses on binary classification of AI-generated images using CNNs with gradient-based explainability (Grad-CAM), not on creating datasets with human-annotated bounding boxes and captions for grounded reasoning. The Cifake dataset contains synthetic images mirroring CIFAR-10 classes but lacks the fine-grained regional annotations with textual descriptions that characterize FakeXplained.

---

6. AI Art Neural Constellation: Revealing the Collective and Contrastive State of AI-Generated and Human Art

URL: View paper

Brief Assessment

AI Art Constellation[57] focuses on contrasting human and AI-generated art through aesthetic and stylistic analysis using art principles, not on detecting synthesis artifacts with grounded annotations for interpretable detection.

---

7. Exploring the naturalness of ai-generated images

URL: View paper

Brief Assessment

Naturalness Exploration[54] focuses on naturalness assessment of AI-generated images through technical and rationality distortions, not on creating a dataset with bounding boxes and captions for synthesis artifact detection and interpretable reasoning.

---

8. RADAR: Reasoning AI-Generated Image Detection for Semantic Fakes

URL: View paper

Brief Assessment

RADAR[27] focuses on semantic fakes (violations of world knowledge/common sense) with the STSF dataset, while FakeXplained targets visual anomalies and AI-generated artifacts across diverse generative models with fine-grained human annotations.

---

9. Wildfake: A large-scale challenging dataset for ai-generated images detection

URL: View paper

Brief Assessment

WildFake[51] focuses on collecting diverse AI-generated images from multiple generators for detection tasks, but does not provide grounded annotations (bounding boxes + captions) for visual artifacts. The datasets serve different purposes: WildFake emphasizes generator diversity and hierarchical structure for cross-generator evaluation, while FakeXplained provides fine-grained human annotations for interpretable detection with spatial grounding.

---

10. Zooming In on Fakes: A Novel Dataset for Localized AI-Generated Image Detection with Forgery Amplification Approach

URL: View paper

Brief Assessment

Zooming In Fakes[53] focuses on localized forgery detection in scene-level edits (sky, ground) using automated semantic calibration, while the original paper provides human-annotated bounding boxes with captions for AI-generated image artifacts. The candidate's BR-GEN dataset addresses different forgery types and uses automated annotation pipelines rather than human expert annotations.

Contribution 2: FakeXplainer detector with progressive training pipeline

Description: An end-to-end system that fine-tunes multi-modal large language models on FakeXplained using a progressive training pipeline integrating supervised fine-tuning and reinforcement learning. The system performs detection, localization, and provides spatially grounded, human-aligned explanations for AI-generated images.

This contribution was assessed against **8 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization

URL: View paper

Brief Assessment

ForgeryGPT[69] focuses on image forgery detection and localization (IFDL) tasks using a three-stage training strategy (image-text alignment, mask-text alignment, task-specific instruction tuning), while the original paper addresses AI-generated image detection with a two-stage progressive training pipeline (SFT followed by GRPO reinforcement learning). The tasks, architectures, and training methodologies differ substantially.

---

2. A collaborative Fusion and Registration Framework for Multi-Modal Image Fusion

URL: View paper

Brief Assessment

Collaborative Fusion Framework[68] addresses multimodal image fusion for infrared/visible images in IoT contexts, not AI-generated image detection or progressive training of MLLMs for artifact localization and explanation.

---

3. Progressive feedback-enhanced transformer for image forgery localization

URL: View paper

Brief Assessment

Progressive Feedback Transformer[71] focuses on image forgery localization using feedback-enhanced transformers for coarse-to-fine detection, not on multi-modal large language models with reinforcement learning for AI-generated image detection and explanation as in the original paper.

---

4. HAMLET-FFD: Hierarchical Adaptive Multi-modal Learning Embeddings Transformation for Face Forgery Detection

URL: View paper

Brief Assessment

HAMLET-FFD[75] focuses on face forgery detection using frozen CLIP parameters with bidirectional cross-modal reasoning, not on progressive training pipelines integrating supervised fine-tuning and reinforcement learning for multi-modal LLMs as in the original paper.

---

5. DA-HFNet: Progressive Fine-Grained Forgery Image Detection and Localization Based on Dual Attention

URL: View paper

Brief Assessment

DA-HFNet[70] focuses on forgery detection in image manipulation (splicing, copy-move) using a hierarchical network for multi-scale feature extraction, not on training multi-modal large language models with reinforcement learning for explainable AI-generated image detection.

---

6. Towards dimension-enriched underwater image quality assessment

URL: View paper

Brief Assessment

Underwater Quality Assessment[72] focuses on underwater image quality assessment via teaching large multimodal models, which is a different domain (underwater imagery) and task (quality assessment) compared to the original paper's AI-generated image detection with progressive RL training.

---

7. Multi-Modal Prompt Learning on Blind Image Quality Assessment

URL: View paper

Brief Assessment

Multi Modal Prompt[73] focuses on blind image quality assessment using multi-modal prompts for CLIP adaptation, not on detecting AI-generated images with progressive RL training for artifact localization and explanation.

---

8. Training-Free In-Context Forensic Chain for Image Manipulation Detection and Localization

URL: View paper

Brief Assessment

Training Free Forensic[74] is a training-free framework for image manipulation localization, while the original paper proposes a progressive training pipeline (SFT + GRPO) for fine-tuning MLLMs on AI-generated image detection. These are fundamentally different approaches—one requires no training, the other is centered on a two-stage fine-tuning methodology.

Contribution 3: State-of-the-art performance with robust explainability

Description: FakeXplainer achieves state-of-the-art detection and localization accuracy while demonstrating strong robustness and out-of-distribution generalization. It uniquely delivers spatially grounded, human-aligned rationales that explain both where and why images appear AI-generated.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Region-Level Data Attribution for Text-to-Image Generative Models

URL: View paper

Brief Assessment

Region Level Attribution[64] focuses on tracing generated image regions back to training data regions for transparency and copyright protection in text-to-image models, not on detecting AI-generated images with explainability. The technical problems are fundamentally different: one is data attribution/provenance tracking, the other is fake image detection with spatial reasoning.

2. CapST: Leveraging Capsule Networks and Temporal Attention for Accurate Model Attribution in Deep-fake Videos

URL: View paper

Brief Assessment

CapST[63] addresses deep-fake video model attribution (identifying which generation model created a video) rather than AI-generated image detection with explainability. The candidate focuses on multiclass classification of deep-fake sources using capsule networks and temporal attention, not on providing spatially grounded explanations for why content appears synthetic.

3. SAGA: Source Attribution of Generative AI Videos

URL: View paper

Brief Assessment

SAGA[66] focuses on source attribution of AI-generated videos (identifying which specific model generated a video), not on detection and localization of AI-generated images with spatially grounded explanations. The tasks, modalities (video vs. image), and objectives (attribution vs. detection with explanation) are fundamentally different.

4. EditTrack: Detecting and Attributing AI-assisted Image Editing

URL: View paper

Brief Assessment

EditTrack[62] addresses a fundamentally different problem: detecting whether a suspicious image was edited from a specific base image and attributing the editing model. FakeXplainer focuses on detecting AI-generated images and explaining why they appear synthetic through spatially grounded reasoning. These are distinct tasks with different technical approaches.

5. Repmix: Representation mixing for robust attribution of synthesized images

URL: View paper

Brief Assessment

RepMix[60] focuses on GAN fingerprinting and attribution (determining which GAN architecture created an image), not on explainable detection with spatially grounded rationales. The candidate addresses robustness to transformations but does not provide human-aligned explanations of synthesis artifacts.

6. Dfda: An analysis of deep learning models to detect deepfake videos

URL: View paper

Brief Assessment

DFDA[61] focuses on deepfake video detection using various deep learning architectures (CNNs, LSTMs, capsule networks) without providing spatially grounded explanations or human-aligned rationales. The paper surveys detection methods but does not address explainability mechanisms like bounding boxes or textual reasoning that characterize FakeXplainer's contribution.

7. Zoom-In to Sort AI-Generated Images Out

URL: View paper

Brief Assessment

Zoom In Sort[67] focuses on a two-stage zoom-in mechanism for detection and localization, achieving 96.39% accuracy with spatial grounding. While both papers address explainability in AI-generated image detection, they employ fundamentally different technical approaches: Zoom In Sort uses a scan-then-zoom iterative refinement process with cropped regions, whereas the original paper uses progressive GRPO training with human-annotated bounding boxes and captions for direct grounding.

8. Interpretable and Reliable Detection of AI-Generated Images via Grounded Reasoning in MLLMs

URL: View paper

Prior Art Analysis

Grounded Reasoning Detection[29] demonstrates that similar prior work exists in achieving state-of-the-art detection and localization accuracy while providing spatially grounded, human-aligned explanations for AI-generated image detection. Both papers construct datasets with bounding box annotations and descriptive captions for synthesis artifacts, fine-tune MLLMs through progressive training strategies, and achieve superior performance in detection, localization, and explainability. The candidate paper's approach of using annotated datasets with visual grounding and multi-stage optimization to achieve robust detection with human-aligned rationales directly parallels the original paper's claimed novelty.

Evidence

Evidence 1 - **Rationale**: Both papers claim to achieve state-of-the-art performance in detection and localization, with the candidate demonstrating similar superior performance claims. - **Original**: extensive experiments show that fakexplainer not only sets a new state-of-the-art in detection and localization accuracy (98.2% accuracy, 36.0% iou), but also demonstrates strong robustness and out-of-distribution generalization, uniquely delivering spatially grounded, human-aligned rationales. - **Candidate**: the resulting model achieves superior performance in both detecting ai-generated images and localizing visual flaws, significantly outperforming baseline methods.

Evidence 2 - **Rationale**: The dataset construction approach is nearly identical, with both papers using bounding boxes and descriptive captions for synthesis artifacts to enable human-aligned grounded reasoning. - **Original**: we constructfakexplained dataset of ai-generated images annotated with bounding boxes and descriptive captions that highlight synthesis artifacts, forming the basis for human-aligned, visually grounded reasoning. - **Candidate**: we construct a dataset of ai-generated images annotated with bounding boxes and descriptive captions that highlight synthesis artifacts, establishing a foundation for human-aligned visual-textual grounded reasoning.

Evidence 3 - **Rationale**: Both papers employ progressive/multi-stage training strategies for MLLMs to achieve the same three objectives: accurate detection, localization, and coherent explanations. - **Original**: leveraging fakexplained, we develop fakexplainer which fine-tunes mllms with a progressive training pipeline, enabling accurate detection, artifact localization, and coherent textual explanations. - **Candidate**: we then finetune mllms through a multi-stage optimization strategy that progressively balances the objectives of accurate detection, visual localization, and coherent textual explanation.

Evidence 4 - **Rationale**: Both papers address the same problem of providing interpretable, robust detection with human-understandable explanations, establishing that this approach existed prior to the original paper's submission. - **Original**: fakexplainer answers "where and why does this image look fake?" with reliable, spatially grounded explanations. extensive experiments show that it achieves state-of-the-art detection accuracy, generalizes well to out-of-distribution images, and remains robust under perturbations while providing huma... - **Candidate**: the rapid advancement of image generation technologies intensifies the demand for interpretable and robust detection methods. although existing approaches often attain high accuracy, they typically operate as black boxes without providing human-understandable justifications.

---

9. DeepGuard: Identification and Attribution of AI-Generated Synthetic Images

URL: View paper

Brief Assessment

DeepGuard[65] focuses on binary detection and model attribution of AI-generated images using ensemble learning methods, achieving high accuracy (98.00%-99.87%). However, it does not address spatially grounded explainability, localization of artifacts, or human-aligned textual reasoning that characterizes FakeXplainer's contribution.

---

10. Unmasking synthetic realities in generative ai: A comprehensive review of adversarially robust deepfake detection systems

URL: View paper

Brief Assessment

Unmasking Synthetic Realities[59] is a systematic review paper that surveys deepfake detection methods across modalities. It does not present a novel detection system with spatially grounded explainability like FakeXplainer, but rather reviews existing approaches without claiming to be the first to propose such methods.

Appendix: Text Similarity Detection

Textual similarity detection checked 31 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Interpretable and Reliable Detection of AI-Generated Images via Grounded Reasoning in MLLMs

Detected in: Core Task (sibling), Contribution: contribution_3

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] FakeXplain: AI-Generated Images Detection via Human-Aligned Grounded Reasoning View paper
- [1] Detecting deepfake images using deep learning techniques and explainable AI methods View paper
- [2] Sida: Social media image deepfake detection, localization and explanation with large multimodal model View paper
- [3] ForenX: Towards Explainable AI-Generated Image Detection with Multimodal Large Language Models View paper
- [4] Towards explainable fake image detection with multi-modal large language models View paper
- [5] Detecting AI-generated images with CNN and Interpretation using Explainable AI View paper
- [6] Funnybirds: A synthetic vision dataset for a part-based analysis of explainable ai methods View paper
- [7] Cifake: Image classification and explainable identification of ai-generated synthetic images View paper
- [8] Detection of AI-generated synthetic images with a lightweight CNN View paper
- [9] MaskSim: Detection of synthetic images by masked spectrum similarity analysis View paper
- [10] AIGI-Holmes: Towards Explainable and Generalizable AI-Generated Image Detection via Multimodal Large Language Models View paper
- [11] Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation View paper
- [12] LayLens: Improving Deepfake Understanding through Simplified Explanations View paper
- [13] Deepfake Image Detection Using Explainable AI and Deep Learning View paper
- [14] DeepExplain: enhancing deepfake detection through transparent and explainable AI model View paper
- [15] Advanced detection of ai-generated images through vision transformers View paper
- [16] Decoding synthetic news: an interpretable multimodal framework for the classification of news articles in a novel news corpus View paper
- [17] Seeing before reasoning: A unified framework for generalizable and explainable fake image detection View paper
- [18] Fakebench: Probing explainable fake image detection via large multimodal models View paper
- [19] Towards Quantitative Evaluation of Explainable AI Methods for Deepfake Detection View paper
- [20] Generation and Detection of Sign Language Deepfakes - A Linguistic and Visual Analysis View paper
- [21] Overcoming diagnostic and data privacy challenges in viral disease detection: an integrated approach using generative AI, vision transformers, explainable AI, and federated learning View paper
- [22] BusterX++: Towards Unified Cross-Modal AI-Generated Content Detection and Explanation with MLLM View paper
- [23] DDL: A Large-Scale Datasets for Deepfake Detection and Localization in Diversified Real-World Scenarios View paper
- [24] Extracting local information from global representations for interpretable deepfake detection View paper
- [25] TruthLens: Explainable DeepFake Detection for Face Manipulated and Fully Synthetic Data View paper
- [26] Explainable deep-fake detection using visual interpretability methods View paper
- [27] RADAR: Reasoning AI-Generated Image Detection for Semantic Fakes View paper
- [28] Enhancing Interpretability in AI-Generated Image Detection with Genetic Programming View paper
- [29] Interpretable and Reliable Detection of AI-Generated Images via Grounded Reasoning in MLLMs View paper
- [30] DDL: A Dataset for Interpretable Deepfake Detection and Localization in Real-World Scenarios View paper
- [31] AiGen-FoodReview: A Multimodal Dataset of Machine-Generated Restaurant Reviews and Images on Social Media View paper
- [32] Explainable AI for Deepfake Detection: A Grad-CAM Approach to Video Forensics View paper
- [33] Advancing AI-Generated Image Detection: Enhanced Accuracy through CNN and Vision Transformer Models with Explainable AI Insights View paper
- [34] YOLOv8 framework for COVID-19 and pneumonia detection using synthetic image augmentation View paper
- [35] Improving AI Generated Image Detection through an Interpretable and Enhanced CNN2D Architecture View paper

- [36] Explainable AI for deepfake detection View paper
- [37] Improving the perturbation-based explanation of deepfake detectors through the use of adversarially-generated samples View paper
- [38] Generative Artificial Intelligence and the Evolving Challenge of Deepfake Detection: A Systematic Analysis View paper
- [39] The Visual Counter Turing Test (VCT2): A Benchmark for Evaluating AI-Generated Image Detection and the Visual AI Index (VAI) View paper
- [40] AI-Generated Image Detection: An Empirical Study and Future Research Directions View paper
- [41] From Prediction to Explanation: Multimodal, Explainable, and Interactive Deepfake Detection Framework for Non-Expert Users View paper
- [42] Performance comparison and visualization of ai-generated-image detection methods View paper
- [43] ThinkFake: Reasoning in Multimodal Large Language Models for AI-Generated Image Detection View paper
- [44] Identity-aware vision-language model for explainable face forgery detection View paper
- [45] METER: Multi-modal Evidence-based Thinking and Explainable Reasoning--Algorithm and Benchmark View paper
- [46] Human-like Content Analysis for Generative AI with Language-Grounded Sparse Encoders View paper
- [47] TruthLens:A Training-Free Paradigm for DeepFake Detection View paper
- [48] The Art of Detection: Methods for Identifying AI-Generated Visual Content View paper
- [49] X2-dfd: A framework for explainable and extendable deepfake detection View paper
- [50] Using multimodal foundation models for detecting fake images on the internet with explanations View paper
- [51] Wildfake: A large-scale challenging dataset for ai-generated images detection View paper
- [52] M3DSYNTH: A dataset of medical 3D images with AI-generated local manipulations View paper
- [53] Zooming In on Fakes: A Novel Dataset for Localized AI-Generated Image Detection with Forgery Amplification Approach View paper
- [54] Exploring the naturalness of ai-generated images View paper
- [55] ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection View paper
- [56] Efficient end-to-end learning for cell segmentation with machine generated weak annotations View paper
- [57] AI Art Neural Constellation: Revealing the Collective and Contrastive State of AI-Generated and Human Art View paper
- [58] GeneVA: A Dataset of Human Annotations for Generative Text to Video Artifacts View paper
- [59] Unmasking synthetic realities in generative ai: A comprehensive review of adversarially robust deepfake detection systems View paper
- [60] Repmix: Representation mixing for robust attribution of synthesized images View paper
- [61] Dfda: An analysis of deep learning models to detect deepfake videos View paper
- [62] EditTrack: Detecting and Attributing AI-assisted Image Editing View paper
- [63] CapST: Leveraging Capsule Networks and Temporal Attention for Accurate Model Attribution in Deep-fake Videos View paper
- [64] Region-Level Data Attribution for Text-to-Image Generative Models View paper
- [65] DeepGuard: Identification and Attribution of AI-Generated Synthetic Images View paper
- [66] SAGA: Source Attribution of Generative AI Videos View paper
- [67] Zoom-In to Sort AI-Generated Images Out View paper
- [68] A collaborative Fusion and Registration Framework for Multi-Modal Image Fusion View paper
- [69] Forgerygpt: Multimodal large language model for explainable image forgery detection and localization View paper
- [70] DA-HFNet: Progressive Fine-Grained Forgery Image Detection and Localization Based on Dual Attention View paper
- [71] Progressive feedback-enhanced transformer for image forgery localization View paper
- [72] Towards dimension-enriched underwater image quality assessment View paper
- [73] Multi-Modal Prompt Learning on Blind Image Quality Assessment View paper
- [74] Training-Free In-Context Forensic Chain for Image Manipulation Detection and Localization View paper
- [75] HAMLET-FFD: Hierarchical Adaptive Multi-modal Learning Embeddings Transformation for Face Forgery Detection View paper