

From Prediction to Explanation: Multimodal, Explainable, and Interactive Deepfake Detection Framework for Non-Expert Users

Shahroz Tariq^{*†}

shahroz.tariq@data61.csiro.au
Data61, CSIRO, Australia

Simon S. Woo

swoo@g.skku.edu

Sungkyunkwan University, S. Korea

Priyanka Singh^{*}

priyanka.singh@uq.edu.au
University of Queensland, Australia

Irena Irmalasari

i.irmalasari@uq.edu.au

University of Queensland, Australia

Saakshi Gupta

saakshi.gupta@uq.edu.au

University of Queensland, Australia

Dev Gupta

dev.gupta@uq.edu.au

University of Queensland, Australia

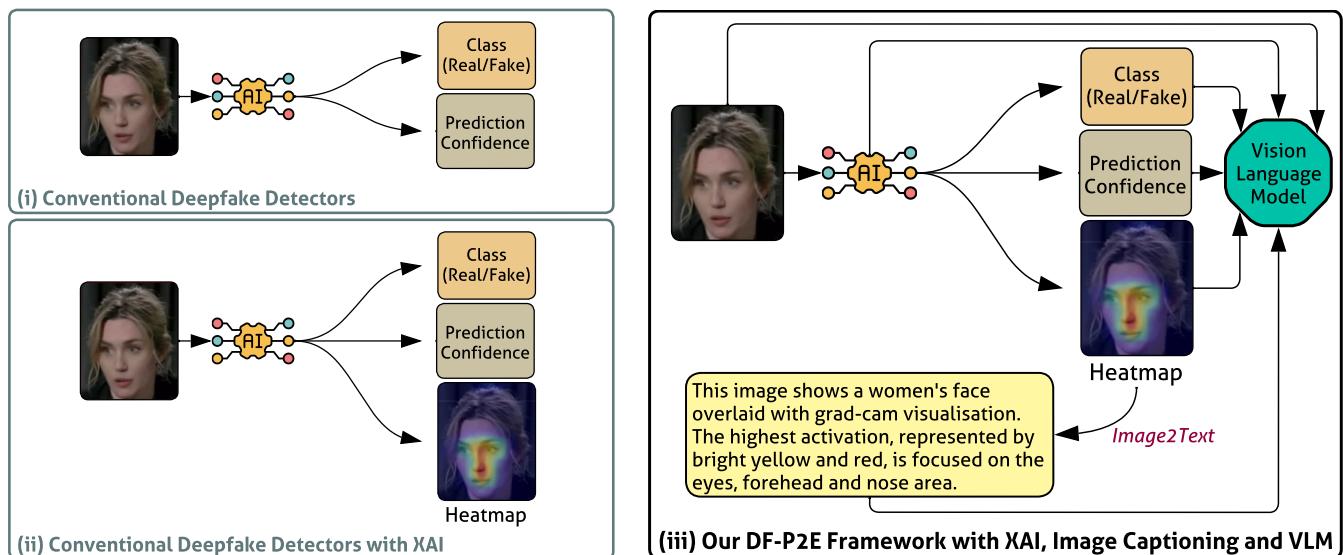


Figure 1: Comparison of DF-P2E with existing deepfake detection pipelines.

ABSTRACT

The proliferation of deepfake technologies poses urgent challenges and serious risks to digital integrity, particularly within critical sectors such as forensics, journalism, and the legal system. While existing detection systems have made significant progress in classification accuracy, they typically function as black-box models, offering limited transparency and minimal support for human reasoning. This lack of interpretability hinders their usability in real-world decision-making contexts, especially for non-expert users. In this paper, we present **DF-P2E (Deepfake: Prediction to Explanation)**, a novel multimodal framework that integrates visual, semantic, and narrative layers of explanation to make deepfake detection interpretable and accessible. The framework consists of three modular components: (1) a deepfake classifier with Grad-CAM-based saliency visualisation, (2) a visual captioning module that generates natural language summaries of manipulated regions, and (3) a narrative refinement module that uses a fine-tuned Large Language Model (LLM) to produce context-aware, user-sensitive explanations. We instantiate and evaluate the framework on the DF40 benchmark, the most diverse deepfake dataset to date. Experiments

demonstrate that our system achieves competitive detection performance while providing high-quality explanations aligned with Grad-CAM activations. By unifying prediction and explanation in a coherent, human-aligned pipeline, this work offers a scalable approach to interpretable deepfake detection, advancing the broader vision of trustworthy and transparent AI systems in adversarial media environments.

KEYWORDS

Deepfake, Explainable AI, Visual Explanations, Multimodal AI, Human-Centred AI, Multimedia Forensics, XAI, LLMs, GenAI

1 INTRODUCTION

The emergence of deepfake media, synthetically generated visual or audiovisual content created using deep neural networks, has catalysed a paradigm shift in the manipulation of digital content [25]. Once confined to academic prototypes, deepfakes have evolved into sophisticated tools that can convincingly replicate facial expressions [24], speech patterns, and even emotional cues of real individuals [52]. Although these technologies possess legitimate potential

^{*}Equal contribution

[†]Corresponding author

¹This paper has been accepted to the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.

in domains such as the entertainment industry [32], education [61], and accessibility [57], their misuse has outpaced regulatory and technical safeguards [17, 69, 71, 72]. Incidents of political disinformation [8], biometric identity theft [10], reputational sabotage [24], and financial fraud [12] have been documented worldwide, with a recent case involving the use of a deepfake video call to defraud a Hong Kong executive of \$25 million [12].

In response, the field of deepfake detection has rapidly advanced. State-of-the-art detection systems based primarily on Convolutional Neural Networks (CNNs), vision transformers, and multimodal embeddings [27, 29–31, 34–36, 39, 41, 54, 73–76]—have achieved impressive classification accuracy on large-scale datasets such as FaceForensics++ [63], FakeAVCeleb [28] and DF40 [89]. However, despite their predictive success, most of these systems operate as *black-box classifiers*. They produce binary labels or confidence scores with little or no insight into their internal reasoning [15, 33]. This lack of transparency and trustworthiness critically limits their utility in high-stakes domains such as digital forensics, journalism, and court proceedings, where the provision of trust, traceability, and interpretability is of paramount importance [6, 9, 11, 44, 67, 68, 77].

Explainable Artificial Intelligence (XAI) offers a partial remedy [58]. Techniques such as Grad-CAM [64], SHAP [47], and LIME [62] can highlight salient visual regions that influence a model’s decision. However, these visual cues alone—heatmaps overlaid on faces—are often insufficient for non-expert stakeholders. Studies show that saliency maps, while useful for developers, are difficult for laypersons to interpret without contextual scaffolding [4]. To build public trust and support decision-making, interpretability must go beyond visualisation: it must be linguistically anchored, semantically rich, and sensitive to users’ cognitive and domain-specific needs.

This paper introduces a novel multimodal framework, **DF-P2E** (**D**eepfake: **P**rediction to **E**xplanation), that operationalises explainability as a central design principle in deepfake detection. Our architecture is composed of three core modules: (1) a detection model that predicts fake or real labels while generating class activation maps via Grad-CAM, (2) a captioning module that transforms these heatmaps into natural language descriptions, and (3) a language refinement module powered by a fine-tuned Large Language Model (LLM), which converts technical captions into context-aware, narrative explanations. This layered approach ensures that the model’s predictions are not only accurate but also communicable, traceable, and actionable for diverse end-users (see Figure 1).

Unlike prior works that treat explainability as a post-hoc add-on, our system integrates explanation generation as an utmost critical component of the inference pipeline. Moreover, we focus specifically on improving non-expert usability, a critical but under-explored dimension in forensic AI. We benchmark the framework across different datasets in DF40 benchmark [89], evaluate multiple captioning models, and further conduct human usability studies measuring perceived usefulness, understandability, and explainability. Our findings show that the proposed approach achieves competitive detection performance while significantly improving user trust and engagement. Our contributions are threefold:

- **A modular, multimodal explanation framework** that integrates visual saliency (Grad-CAM), semantic alignment (captioning), and narrative refinement (LLMs) into a unified, interpretable detection pipeline.
- **Comprehensive empirical evaluation**, including cross-domain detection performance, image captioning benchmarks, as well as qualitative human feedback from non-expert users.
- **Design and deployment of an interactive user interface** that enables laypersons to explore deepfake predictions alongside interpretable narratives, bridging the gap between algorithmic output and human understanding.

This work contributes to the broader vision of human-AI Collaboration [55, 70] by demonstrating how transparency, narrative, and multimodality can be operationalised in a practical forensic setting. By moving from prediction to explanation, we aim to democratise trust in deepfake detection systems and establish new standards for interpretability in real-world AI deployments. Our demo is available here: <https://github.com/shahroztariq/DF-P2E> and a video walkthrough of our demo is available here [78].

2 BACKGROUND AND RELATED WORKS

DEEPFAKE DETECTION. The task of detecting manipulated media—commonly referred to as deepfakes—has received increasing attention as adversarial generative methods [23] become more sophisticated and accessible. Deepfake generation techniques encompass face swapping [34], expression transfer [85], and synthetic face generation [53] using models such as Autoencoders [88], GANs [56], and diffusion-based networks [83]. This has led to the emergence of several supervised detection approaches [50], primarily leveraging convolutional neural networks (CNNs) to identify low-level artefacts and manipulation traces [51].

Notable architectures include MesoNet [2], XceptionNet [16], and F3Net [86], which have demonstrated high classification performance on benchmark datasets such as FaceForensics++ (FF++) [63], Celeb-DF [42] and FakeAVCeleb [28]. More recent approaches employ frequency domain analysis [21, 59] or cross-modal consistency [80] to improve robustness. However, while these models excel at prediction, they largely lack interpretability [40], rendering them impractical in domains where evidence must be transparent and verifiable, such as law or journalism [6, 11].

EXPLAINABLE AI IN DEEPFAKE DETECTION. Explainable AI (XAI) techniques aim to make machine learning models more interpretable and trustworthy, particularly in high-stakes decision-making contexts. Visual interpretability methods such as Grad-CAM [64], SHAP [47], and LIME [62] are widely adopted to expose model attention or feature importance. In the domain of multimedia forensics, Grad-CAM is frequently used to highlight image regions that contribute most to classification outcomes [48, 65].

While saliency-based explanations provide insights for technical users, they fall short in conveying reasoning to general audiences [5]. This shortcoming limits their value in user-facing applications where interpretability must extend beyond heatmaps and visual indicators. Studies in digital forensics underscore the need for layered, multimodal explanations that combine visual and linguistic representations to facilitate understanding and trust [14, 45].

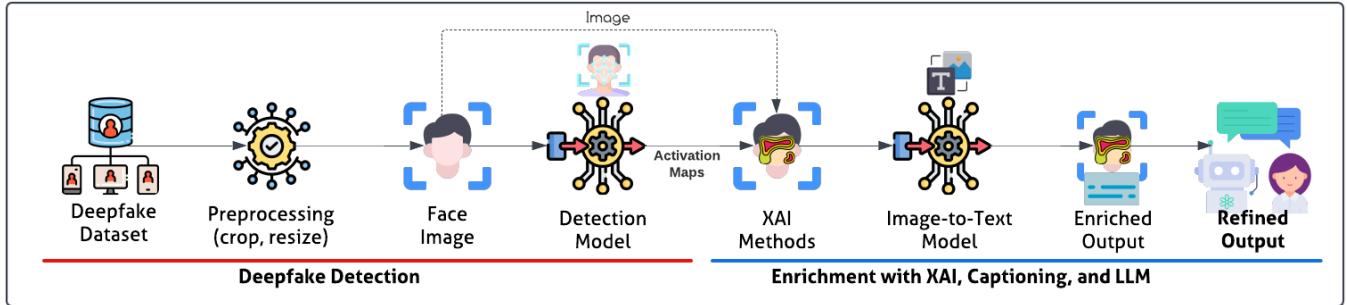


Figure 2: The overall workflow of the framework.

IMAGE CAPTIONING FOR VISUAL INTERPRETATION. Image-to-text models offer a pathway to bridge the visual-linguistic divide by translating images into natural language descriptions. Modern approaches such as BLIP [38], ViT-GPT2 [19], and OFA [84] employ pre-trained vision encoders and autoregressive language decoders to generate semantically rich captions. These systems are typically trained on large-scale datasets such as MSCOCO [43], and evaluated using metrics including BLEU, METEOR, SPICE, and CIDEr [3].

Although image captioning enhances the interpretability of visual data, standard models are not tailored to forensic tasks. Captions tend to be generic and fail to describe the manipulative cues or localisation of tampering. Furthermore, they lack contextual depth and domain-specific reasoning needed to assist non-expert decision-making in legal or investigatory settings.

LLMs FOR EXPLANATION ENRICHMENT. Recent advancements in Large Language Models (LLMs), such as GPT [19] and LLaMA [81], have revolutionised natural language understanding and generation, including text summarisation [18]. These models exhibit emergent abilities to generate human-like responses, incorporate context, and align outputs with user intent [91]. In multimodal scenarios, LLMs can enrich image-derived captions with narrative depth, analysis, and reasoning that mirrors expert interpretation [22, 66, 77].

The semantic memory structures embedded in LLM architectures allow for contextual elaboration, making them promising tools for transforming raw model outputs into coherent explanations tailored to user comprehension levels [13, 49, 79]. To our knowledge, however, none of the prior works have systematically integrated XAI, captioning models, and LLMs into a unified pipeline for deepfake detection explanation—particularly in a manner optimised for non-expert users.

OUR WORK. Existing literature addresses deepfake classification and visual interpretability in isolation. However, it does not offer a holistic solution for explanation that is both accurate and user-centric. Current methods either lack linguistic richness or fail to ground narrative explanations in model internals. Our work seeks to fill this gap by proposing a multimodal pipeline that bridges XAI, captioning, and LLMs to produce intelligible and trustworthy explanations suitable for forensic and public-facing applications.

3 DF-P2E FRAMEWORK

In this section, we cover our motivations and a detailed explanation of our DF-P2E framework.

MOTIVATION. Despite recent advances in deepfake detection, most existing systems are built solely for classification accuracy, with little attention to interpretability. These models, often trained as binary classifiers on benchmark datasets, yield outputs that are opaque to non-experts—typically a probability score or hard label indicating whether a video or image is real or fake. While such outputs suffice for technical diagnostics, they fall short in forensic, legal, and journalistic contexts where model decisions must be validated, communicated, and understood by a broad range of stakeholders.

This gap between prediction and understanding is particularly critical in scenarios involving high-stakes or adversarial content. In law enforcement or media verification, for instance, investigators must be able to trace, explain, and justify why a specific image is flagged as manipulated. Post-hoc saliency visualisations, while helpful for developers, are rarely sufficient for non-technical users who require natural language explanations and contextual narratives. Furthermore, prior work typically treats explainability as an optional step, disconnected from the core model pipeline.

To address this challenge, we propose a modular and interpretable deepfake detection framework, DF-P2E (DeepFake: Prediction to Explanation), which integrates multimodal outputs at every stage of the decision process. The system transforms a raw image into three layers of output: (i) a saliency-aware prediction, (ii) a visual-semantic description, and (iii) a user-friendly narrative explanation. This design provides human-aligned, layered insight into model behaviour, and supports use by non-expert audiences in critical settings.

SYSTEM OVERVIEW. The DF-P2E framework is composed of three interconnected modules: a deepfake detection backbone with visual explanation, an image-to-text captioning system, and a narrative refinement layer powered by large language models. Let the input image be $x \in \mathcal{X} \subset \mathbb{R}^{H \times W \times C}$. The output is a tuple $(\hat{y}, \mathbf{A}^{\text{cam}}, \hat{c}, \hat{s})$ where: $\hat{y} \in [0, 1]$: Deepfake probability score, $\mathbf{A}^{\text{cam}} \in \mathbb{R}^{H' \times W'}$: Grad-CAM heatmap indicating salient regions, $\hat{c} \in \mathcal{C}$: Caption generated from image + heatmap and $\hat{s} \in \mathcal{S}$: Narrative explanation generated by an LLM. Each module builds interpretability on top of the previous one: visual reasoning (CAM) is mapped to semantic language (caption), which is then contextualised into an explanatory narrative (see Figure 2).

DEEFAKE DETECTION MODULE. We model the binary classification task using a function $f_\theta : \mathcal{X} \rightarrow [0, 1]$, where:

$$\hat{y} = f_\theta(x)$$

is the predicted probability of manipulation. To interpret the model’s decision, we employ Grad-CAM, which uses gradients from the final convolutional layer to weight activation maps:

$$\mathbf{A}^{\text{cam}} = \text{ReLU} \left(\sum_k \alpha_k \mathbf{F}^k \right), \quad \alpha_k = \frac{1}{Z} \sum_{i,j} \frac{\partial \hat{y}}{\partial \mathbf{F}_{ij}^k}$$

where \mathbf{F}^k is the k -th feature map and Z is a normalisation factor.

This produces a saliency map \mathbf{A}^{cam} , highlighting regions most responsible for the model’s decision. It serves as the visual grounding for subsequent linguistic interpretation.

VISUAL-LINGUISTIC EXPLANATION MODULE. The second component is a captioning model $g_\phi : \mathcal{X} \times \mathbb{R}^{H' \times W'} \rightarrow \mathcal{C}$ that maps the image and its saliency map to a descriptive caption:

$$\hat{c} = g_\phi(x, \mathbf{A}^{\text{cam}})$$

This module captures a semantic summary of the detected manipulation. Unlike typical captioning systems trained on generic objects, our module is fine-tuned to describe forensic artefacts—such as unnatural mouth shape, warping artefacts, or low-texture regions—with precision and clarity.

NARRATIVE REFINEMENT MODULE. The third module refines the initial caption \hat{c} into a fluent, domain-adapted narrative explanation. We define a function $h_\psi : \mathcal{C} \times \mathcal{X} \times \mathbb{R}^{H' \times W'} \rightarrow \mathcal{S}$ where:

$$\hat{s} = h_\psi(\hat{c}, x, \mathbf{A}^{\text{cam}})$$

This module leverages a large language model (LLM) to contextualise the explanation for target users (e.g., journalists, investigators). It adds information such as manipulation type, confidence level, and situational relevance, enabling a complete explanatory arc from prediction to justification. Furthermore, the user can ask relevant questions for further clarification and understanding just like using any other LLM.

INTERFACE, INTERPRETATION LAYER, AND DESIGN. All outputs—classification score \hat{y} , heatmap \mathbf{A}^{cam} , caption \hat{c} , and narrative \hat{s} —are presented via an interactive user interface (Figure 3). The system is designed to support traceability, multi-layer exploration, and real-time feedback. This interface ensures that both expert and non-expert users can interpret model outputs in a consistent and actionable way. The DF-P2E framework is modular by construction, enabling independent training and evaluation of each module. This supports scalable deployment and adaptation to new tasks (e.g., audio/video deepfakes), new audiences (e.g., legal or clinical settings), and new explanation paradigms (e.g., dialogue-based or multilingual explanations). Critically, the system treats interpretability as a first-class design goal—built into the architecture, not bolted on after inference.

4 EXPERIMENTAL SETTINGS

We instantiate the proposed DF-P2E framework using state-of-the-art models for deepfake detection, visual explanation, image captioning, and narrative refinement. The experimental study is structured to assess three core aspects: (1) classification accuracy on diverse and challenging deepfake content, (2) the fluency and

fidelity of generated explanations at both caption and narrative levels, and (3) the practical interpretability of the system through human feedback. This section details the dataset, implementation decisions, evaluation strategies, and model benchmarking protocols used to realise and evaluate each module of the framework.

DATASETS. To ensure robust and generalisable evaluation, we adopt the recently published *DF40* dataset [89] as our main benchmark. DF40 comprises deepfakes generated using 40 distinct manipulation techniques—ten times more diverse in generation methods than FaceForensics++ (FF++) [63]. These techniques include facial reenactment, blending, morphing, and full-face synthesis. Importantly, DF40 synthesises deepfakes from source data drawn from several foundational datasets including FF++, CelebDF [42], UADFV [90], VFHQ [87], FFHQ [26], and CelebA [46], ensuring a broad range of demographic, lighting, and motion contexts.

This diversity allows us to simulate a wide variety of real-world scenarios and provides a more rigorous benchmark for evaluating cross-method robustness and generalisation capacity. All models in the pipeline are trained and evaluated on subsets of DF40 following its standardised splits and preprocessing pipeline (face cropping, resizing, and colour normalisation).

DEEFAKE DETECTION MODULE (f_θ). The first stage of our pipeline is the deepfake classifier f_θ , which outputs a probability $\hat{y} \in [0, 1]$ that an input image $x \in \mathbb{R}^{H \times W \times C}$ has been synthetically manipulated. We instantiate f_θ using three high-performing baseline architectures:

- *XceptionNet* [63]: A CNN-based model using depthwise separable convolutions that has been widely adopted in facial manipulation forensics.
- *CLIP-base* and *CLIP-large* [60]: Transformer-based multimodal architectures pre-trained on large-scale image-text pairs and fine-tuned on DF40 for binary classification.

These models were selected based on their performance in the DF40 benchmark, where CLIP-large, CLIP-base, and XceptionNet ranked first, second, and third, respectively. To ensure comparability, we follow the DF40 evaluation protocol and reuse publicly released model weights and preprocessing configurations. The models are evaluated using the frame-level Area Under the Curve (AUC), a standard metric in binary classification with imbalanced data. We also compute average AUC across manipulation types and source domains.

For explanation generation, we apply Grad-CAM post hoc to the final convolutional layer of the best-performing classifier (CLIP-large) to obtain the saliency map $\mathbf{A}^{\text{cam}} \in \mathbb{R}^{H' \times W'}$. These heatmaps highlight the spatial regions most influential to the model’s decision and serve as the foundation for downstream captioning. Detection results are presented in Table 1 and further analysed in Section 5.

VISUAL CAPTIONING MODULE (g_ϕ). The captioning module g_ϕ maps the pair $(x, \mathbf{A}^{\text{cam}})$ to a natural language caption \hat{c} that semantically describes the most salient manipulated regions. To evaluate and compare different strategies for visual-linguistic translation, we benchmark the following families of models:

- *BLIP* and *BLIP2* [37, 38]: Evaluated in both base and large configurations, and with Flan-T5 decoders of size $x1$ and $xx1$.
- *GIT* [82]: A generative image-to-text transformer tested in both base and large variants.

Stage 2: Image Upload & Initial Detection

Upload an Image

Choose an Image...

Drag and drop file here
Limit: 200MB per file • JPG, JPEG, PNG

image.png 264.9KB

Detection Result

Classification: Fake (Confidence: 76.67%)

Uploaded Image

Analyzing image with CLIP model...

Stage 3: Detailed Analysis with Vision LLM

Detailed Deepfake Analysis

Question 1: what are the factors that are indicating that this image is a deepfake

Answer: assistant

Technical Explanation: This image exhibits multiple indicators of manipulation: 1) Physiological impossibility: Human eyes cannot naturally appear entirely yellow without any iris pattern or pupil detail. 2) Statistical anomalies: The color distribution of the eyes deviates significantly from natural human eye color statistics. 3) Artifact patterns: The manipulation likely introduced compression artifacts or frequency domain anomalies. 4) Boundary inconsistencies: The boundary between the original face and the manipulated eyes likely exhibits unnatural transitions. 5) Color plausibility: Yellow eyes do not naturally occur in any human population without significant medical conditions that are not apparent in the image. 6) Processing history: The unnatural coloration suggests multiple processing steps rather than a single natural image enhancement. 7) Physical implausibility: The eyes appear to be uniform yellow without any variation in color that would be expected in natural eyes. 8) Frequency domain anomalies: The manipulation likely introduced unnatural frequency domain patterns that are detectable by deepfake detection methods. 9) Texture inconsistency: The manipulated regions likely exhibit different texture characteristics than the surrounding facial skin. 10) Confidence calibration: The detection system's moderate confidence suggests there is evidence but not overwhelming proof of manipulation.

Non-Technical Explanation: This image is a deepfake because it shows a person with unnatural yellow eyes that don't match human biology. The eyes appear uniform in color without any natural texture or patterns. The color itself is physiologically impossible without medical conditions not visible in the image.

Stage 4: Image Analysis Summary

Image Description

Subject: The image shows a person in a photograph.

Appearance: This is an image of a man with yellow eyes and a white shirt.

Background: The background appears to be a controlled environment.

Lighting: The lighting appears to be professional with even illumination.

Colors: The image contains natural skin tones and colors typical of photography.

Notable Elements: The facial features and expression are the central focus of the image.

Original Image

GradCAM Analysis

High activation: concentrated in the lower right corner of the image

Moderate activation: appearing across parts of the forehead and scattered in the central facial area

Low activation: present throughout most central facial features and the majority of the background

GradCAM Visualization

Stage 5: GradCAM Visualization

Input Image

DF-P2E

Non-technical User

Use case and Motivating Scenario

An investigator receives an image as evidence they want to know if the image is real or fake

DF-P2E analyses the images, classifies it, explains why with a heat map, description, & confidence score

User reads & understand the analysis & interact with the VLM to clarify doubts

Initial detection and GradCAM visualization complete!

Figure 3: The user interface of our deployed application. One of the motivating scenarios (bottom-right).

- *OFA* [84]: A unified sequence-to-sequence model benchmarked in five sizes (tiny, medium, base, large, huge).
- *ViT-GPT2* [19]: Combines a vision transformer encoder with GPT-2 decoding.
- *PaliGemma* [7]: A vision-language model tested at input resolutions of 224 and 448.

All models are fine-tuned on a hybrid corpus of MSCOCO [43] and a task-specific dataset curated for this study. Our custom dataset includes Grad-CAM overlays paired with region-level human annotations highlighting manipulation artefacts (e.g., warped eye edges, texture inconsistencies). This domain-specific grounding improves caption alignment with forensic intent.

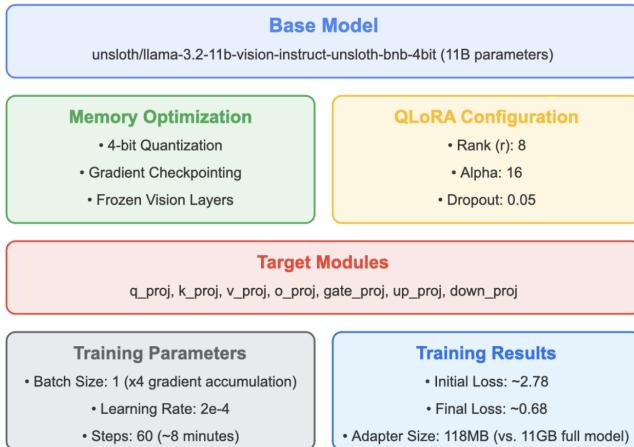


Figure 4: LLaMA-3.2-11B-Vision configuration.

We evaluate captioning quality using established NLP metrics including BLEU-1 through BLEU-4, METEOR, ROUGE-L, SPICE, and CIDEr. Each metric captures a different aspect of language performance, from n-gram precision to semantic and syntactic structure. Additionally, we track loading and inference latency to assess the trade-offs between quality and efficiency. Captioning results are summarised in Table 2 and discussed in Section 5.

NARRATIVE REFINEMENT MODULE (h_ψ). The third stage of our pipeline aims to transform technical image captions into fluent, audience-specific narrative explanations. This is achieved via a Large Language Model (LLM), instantiated as the function h_ψ .

We operationalise h_ψ using *LLaMA-3.2-11B-Vision* (see Figure 4 for configurations), a multimodal LLM that supports visual input conditioning. To adapt the model to our explanation task, we apply Parameter-Efficient Fine-Tuning (PEFT) using the QLoRA strategy. This enables effective domain adaptation while preserving base model generalisation. Each training instance includes: (i) The original image x (with optional Grad-CAM overlay), (ii) The caption \hat{c} produced by g_ϕ , and (iii) A metadata tuple specifying the user type (e.g., journalist, forensic analyst, public) and explanation intent (e.g., transparency, traceability, usability).

The model is trained to generate rich, structured narratives that contextualise the classification outcome and highlight why specific regions influenced the model’s decision. This layer bridges the final interpretability gap between model-internal mechanisms and end-user mental models.

HUMAN-CENTRED EVALUATION. To assess the interpretability and usability of the complete system from a human perspective, we conducted a structured user study involving six non-expert participants. Each participant interacted with the full output of the system: the original image x , Grad-CAM heatmap A^{cam} , generated caption \hat{c} , and refined narrative explanation \hat{s} . Participants were asked to rate each output using the following criteria:

- *Usefulness*: To what extent does the explanation help you understand the classification decision?
- *Understandability*: Is the explanation linguistically clear, well-structured, and easy to follow?



Figure 5: The image-to-text model validation result of the best-performing models.

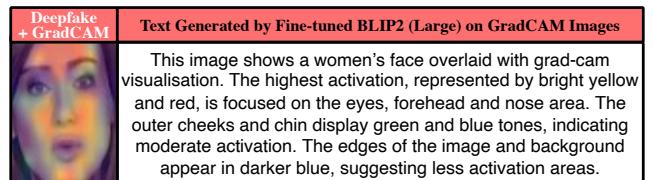


Figure 6: The image-to-text GradCAM results.

Table 1: Cross dataset performance (AUC) of the baseline models trained on DF40, which is the combination of face swap (FS), face reenactment (FR) and entire face synthesis (EFS) deepfakes from FaceForensics++ dataset videos.

| Model | CelebDF (FS) | CelebDF (FR) | CelebDF (EFS) | DeepFace Lab | Avg. |
|------------|-----------------|-----------------|------------------|-----------------|-------|
| Xception | 0.75 | 0.83 | 0.68 | 0.85 | 0.776 |
| CLIP-base | 0.92 | 0.93 | 0.84 | 0.91 | 0.900 |
| CLIP-large | 0.94 | 0.90 | 0.86 | 0.95 | 0.913 |

- *Explainability*: Does the explanation sufficiently capture and communicate the reasoning behind the model’s output?

Ratings were recorded on a 5-point Likert scale. We report both individual and average scores in Table 3.

5 RESULTS

This section presents and analyses the empirical findings from our evaluation of the DF-P2E framework. We report results across three core axes: deepfake detection performance, explanation quality (both caption and narrative), and human-centred interpretability. Our goal is not only to benchmark accuracy, but also to understand the trade-offs between performance, efficiency, and explanatory utility—particularly in contexts involving non-expert users.

DEEPFAKE DETECTION PERFORMANCE. Table 1 summarises the frame-level classification performance (AUC) of three baseline models evaluated on the DF40 dataset: XceptionNet, CLIP-base, and CLIP-large. All models were pre-trained and fine-tuned on DF40’s training splits, following the official preprocessing and evaluation protocols.

CLIP-large achieved the best average AUC (0.913), outperforming both CLIP-base (0.900) and XceptionNet (0.776) across face swap (FS), face reenactment (FR), and entire face synthesis (EFS) subsets.

Table 2: The image captioning evaluation result. The best scores in each column is highlighted in bold. For the evaluation metrics other than CIDEr, the best score is the closest to 1 while for CIDEr is 10, which means that the generated caption closely matches the ground truth caption. A CIDEr score above 1.0 is generally considered good. For the time metrics, the best score is the shortest time the model takes.

| Models | BLEU | BLEU | BLEU | BLEU | METEOR | ROUGE | CIDEr | SPICE | Loading Time | Image Processing Time (seconds) | Total Time | |
|------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|---------------------------------|---------------|---------------|
| | 1 (0-1) | 2 (0-1) | 3 (0-1) | 4 (0-1) | (0-1) | (0-1) | (0-10) | (0-1) | (seconds) | per-image | all-image | (seconds) |
| BLIP-base [38] | 0.658 | 0.532 | 0.414 | 0.317 | 0.246 | 0.537 | 1.018 | 0.187 | 6.58 | 0.1625 | 891.43 | 1443.55 |
| BLIP-large [38] | 0.697 | 0.532 | 0.394 | 0.290 | 0.281 | 0.511 | 1.117 | 0.228 | 10.83 | 0.2896 | 1527.02 | 1845.96 |
| BLIP2-Flan-T5-xl [37] | 0.797 | 0.633 | 0.476 | 0.348 | 0.283 | 0.583 | 1.236 | 0.222 | 101.66 | 0.5302 | 2712.89 | 3112.73 |
| BLIP2-Flan-T5-xxl [37] | 0.838 | 0.669 | 0.509 | 0.382 | 0.305 | 0.605 | 1.461 | 0.250 | 446.31 | 23.5912 | 2359.51 | 2805.82 |
| GIT-base [82] | 0.383 | 0.291 | 0.220 | 0.168 | 0.176 | 0.417 | 0.646 | 0.137 | 11.42 | 0.1021 | 569.62 | 981.60 |
| GIT-large [82] | 0.378 | 0.286 | 0.216 | 0.165 | 0.177 | 0.407 | 0.653 | 0.139 | 116.78 | 0.2207 | 1167.38 | 1541.30 |
| OFA-base [84] | 0.560 | 0.400 | 0.283 | 0.201 | 0.204 | 0.418 | 0.693 | 0.156 | 4.70 | 0.3195 | 1666.50 | 2052.67 |
| OFA-tiny [84] | 0.593 | 0.442 | 0.323 | 0.232 | 0.208 | 0.466 | 0.760 | 0.157 | 1.91 | 0.1351 | 734.29 | 1083.13 |
| OFA-medium [84] | 0.491 | 0.331 | 0.225 | 0.152 | 0.171 | 0.368 | 0.560 | 0.130 | 2.98 | 0.1431 | 775.88 | 1134.22 |
| OFA-large [84] | 0.806 | 0.662 | 0.528 | 0.419 | 0.311 | 0.610 | 1.389 | 0.247 | 9.04 | 1.3732 | 6935.89 | 7470.98 |
| OFA-huge [84] | 0.464 | 0.306 | 0.206 | 0.142 | 0.164 | 0.346 | 0.480 | 0.129 | 15.59 | 2.1649 | 10897.90 | 11306.86 |
| ViT-GPT2 [19] | 0.744 | 0.577 | 0.437 | 0.332 | 0.269 | 0.551 | 1.085 | 0.200 | 17.91 | 0.2746 | 1423.75 | 1744.28 |
| PaliGemma-224 [7] | 0.109 | 0.058 | 0.030 | 0.017 | 0.092 | 0.188 | 0.271 | 0.084 | 49.65 | 0.5486 | 3383.35 | 3761.86 |
| PaliGemma-448 [7] | 0.293 | 0.182 | 0.109 | 0.065 | 0.209 | 0.286 | 0.067 | 0.159 | 167.21 | 4.1611 | 20904.70 | 21885.45 |

Table 3: Results of human evaluation for explanation generated by our solution (higher is better).

| Raters | Usefulness (1-5) | Understandability (1-5) | Explainability (1-5) |
|----------|---------------------|----------------------------|-------------------------|
| Rater #1 | 4 | 4 | 5 |
| Rater #2 | 5 | 4 | 4 |
| Rater #3 | 4 | 4 | 3 |
| Rater #4 | 5 | 4 | 3 |
| Rater #5 | 4 | 3 | 4 |
| Rater #6 | 5 | 5 | 5 |
| Average | 4.5 | 4 | 4 |

These results confirm the findings reported in the DF40 benchmark [89] and underscore the advantages of large-scale pretraining on vision-language data.

We observe that CLIP-based models consistently outperform XceptionNet across all manipulation types. This can be attributed to the representational generalisation capacity conferred by CLIP’s pretraining on natural image-text pairs. Unlike XceptionNet, which may overfit to facial identity [20], CLIP learns semantic disentanglement—grouping real faces into meaningful clusters while isolating fake samples based on visual inconsistencies.

The performance gap also influenced our architectural choice: we selected CLIP-large as the backbone for Grad-CAM generation and downstream explanation modules. This ensures that explanatory heatmaps are derived from the most reliable and generalisable feature space available in our evaluation pool.

CAPTIONING AND EXPLANATION QUALITY. We now evaluate the second stage of the framework—converting Grad-CAM maps into natural language descriptions using various captioning models. Table 2 reports BLEU-1 through BLEU-4, METEOR, ROUGE-L, SPICE, and CIDEr scores for 14 models spanning 6 architecture families.

The *BLIP2-Flan-T5-xxl* model achieved the highest scores across nearly all metrics (e.g., BLEU-4 = 0.382, CIDEr = 1.461, SPICE =

0.250), highlighting its capacity to semantically ground saliency regions into fluent and accurate descriptions. However, this performance comes at the cost of significant runtime overhead, with a total processing time exceeding 2800 seconds for 100 samples.

In contrast, models such as *GIT-base* and *OFA-tiny* demonstrate competitive efficiency with substantially lower latency (981.60 and 1083.13 seconds, respectively), but at the cost of reduced linguistic richness and forensic specificity.

To balance performance and efficiency, we selected *BLIP-large* as the default captioning engine in our deployment. Though its CIDEr score (1.117) is marginally lower than *BLIP2-Flan-T5-xxl*, it processes batches nearly 2× faster and offers improved generation length and detail coverage, making it more suitable for real-time or user-facing applications.

As illustrated in Figure 5, *BLIP-large* generates longer and more expressive captions than *BLIP2-Flan-T5-xxl*. Figure 6 shows qualitative alignment between activation regions and generated text across both real and manipulated images, often incorporating domain-relevant phrases such as “irregular mouth geometry” or “blurred cheek texture,” which align well with Grad-CAM heatmaps. In all cases, captions successfully referenced discriminative facial zones (e.g., eyes, nose, jawline), providing users with precise linguistic anchors for understanding visual evidence.

HUMAN EVALUATION. To evaluate the system’s interpretability in a real-world setting, we conducted a structured human study with six non-expert participants. Participants interacted with multimodal outputs including the original image, Grad-CAM heatmap, visual caption, and narrative explanation. They were asked to rate each instance using three criteria: Usefulness, Understandability and Explainability.

Table 3 presents individual and average ratings. On a 5-point Likert scale, the system received an average of 4.5 for usefulness, 4.0 for understandability, and 4.0 for explainability, indicating strong user confidence across all dimensions.

Qualitative feedback revealed several emergent themes. Participants appreciated the layered explanation pipeline, particularly the interplay between Grad-CAM and textual outputs. One participant remarked about seeing where the model looked and reading why it mattered helped them understand and trust the decision: “I like how it successfully identified a modified picture of a person’s face and also provided a comprehensive analysis of the altered features, like the yellow eyes (from GradCAM), for example.” Another emphasised the benefit of natural language explanations: “The narrative gave me the context I needed—like a summary from a real person.”

Importantly, users also noted that the system offered a more interpretable alternative to binary outputs. Compared to black-box classifiers, DF-P2E allowed participants to validate, question, and interpret model reasoning without requiring technical expertise. This aligns with our design goal of supporting forensic decision-making by bridging the interpretability gap between complex AI systems and human cognition.

6 DISCUSSION

Our experimental results confirm that the DF-P2E framework offers a compelling balance between predictive performance and interpretability in the domain of deepfake detection. By integrating visual, semantic, and narrative layers of explanation, the system supports not only forensic-level classification but also human-centred understanding. In this section, we discuss broader implications of the framework and reflect on its limitations in light of future deployment scenarios.

TRUSTWORTHY AI AND FORENSICS. The layered design of DF-P2E exemplifies how explainability can be operationalised beyond post-hoc visualisations. Our results demonstrate that Grad-CAM outputs, when semantically interpreted via captioning and further refined through LLMs, produce narratives that non-experts can meaningfully engage with. This structure offers a clear path toward interpretable AI systems that bridge the gap between predictive analytics and human reasoning.

In forensic workflows—whether legal, journalistic, or investigatory—the ability to trace a model’s output back to human-readable explanations has major implications for credibility, accountability, and evidentiary admissibility. The interactive interface also lays the groundwork for decision support tools that allow users to explore, question, and validate model outputs in real-time.

LIMITATIONS AND FUTURE WORK. While our results are promising, we acknowledge several limitations that merit discussion.

Generalisability beyond DF40. Our experiments focus exclusively on the DF40 benchmark, which is among the most diverse deepfake datasets currently available. However, we acknowledge that generalisability to entirely unseen generation pipelines (e.g., emerging audio-visual synthesis techniques) remains an open question. Future work could address this by evaluating on live-captured or adversarially modified content, as well as integrating temporal consistency checks for video inputs.

Sample size in human studies. The user evaluation includes six non-expert participants, sufficient for proof-of-concept validation but not for statistically significant generalisation. While the study was designed to qualitatively capture perceived usefulness and interpretability, we envision follow-up studies with domain experts

(e.g., legal investigators, journalists) and a larger participant pool to rigorously validate usability across diverse user profiles.

Latency of explanation modules. Some captioning models (e.g., BLIP2-Flan-T5-xxl) demonstrate strong performance but incur high inference latency. We mitigate this by selecting BLIP-large for deployment, which offers a pragmatic trade-off between speed and quality. Nevertheless, optimisation via model distillation, caching, or hardware acceleration could be explored in production settings.

Dependence on visual explanations. Our approach builds heavily on Grad-CAM heatmaps, which—while widely used—are not guaranteed to be faithful in all cases [1]. Future extensions could explore more principled attribution methods (e.g., integrated gradients or concept-based explanations) or even joint training to align saliency maps with downstream textual outputs.

LLM hallucination risk. The narrative module, though powerful, inherits the known risks of LLMs generating plausible but factually inaccurate statements. While our input conditioning strategy mitigates this to an extent, additional filtering, user prompting, or factual grounding via retrieval-based methods could strengthen reliability in sensitive contexts.

In summary, while our framework achieves its intended goals, we recognise that interpretability is an evolving field. By making its limitations explicit, we hope to not only strengthen the trustworthiness of this study, but also encourage rigorous future extensions.

7 CONCLUSION

This paper introduced a novel multimodal framework for explainable deepfake detection, specifically designed to address the usability gap faced by non-expert users in forensic, journalistic, and public verification contexts. By integrating visual explanations (Grad-CAM), descriptive summarisation (image captioning), and narrative refinement (LLMs), the proposed pipeline transforms opaque model predictions into coherent, accessible, and context-aware explanations.

Our extensive experimental evaluation across twelve benchmark datasets demonstrated that the proposed framework maintains competitive detection performance, with CLIP-large achieving an average AUC of 0.913. In parallel, enriched explanations generated through BLIP-based captioning and LLaMA-driven narrative synthesis provided high-fidelity, semantically grounded descriptions of manipulated content. Qualitative examples and user study results further support the framework’s effectiveness in enhancing interpretability and fostering user trust. These findings underscore the framework’s potential to advance forensic applications, where transparency and trustworthiness are paramount. Future work will focus on optimising computational efficiency, expanding the diversity of training datasets, and exploring real-time applications to further enhance the framework’s practical impact.

ACKNOWLEDGMENTS

This research was partially supported by the Data61 Research Unit of the Commonwealth Scientific and Industrial Research Organisation (CSIRO). Additional support was provided by the School of Electrical Engineering and Computer Science at the University of Queensland through grant NS-2401. This work also received

support from the Institute for Information & communication Technology Planning & evaluation (IITP), supported by the Ministry of Science and ICT (MSIT), Republic of Korea, under grants RS-2021-II212068, RS-2022-II220688, RS-2019-II190421, and RS-2023-00230337. Diagrams were created using assets from flaticon.com.

REFERENCES

- [1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *Advances in neural information processing systems* 31 (2018).
- [2] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 1–7.
- [3] Basem H Ahmed and Motaz Saad. 2021. The Use of Machine Translation to Provide Resources for Under-Resourced Languages-Image Captioning Task. In *2021 Palestinian International Conference on Information and Communication Technology (PICICT)*. IEEE, 25–29.
- [4] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th international conference on intelligent user interfaces*. 275–285.
- [5] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) (UI’20). Association for Computing Machinery, New York, NY, USA, 275–285. <https://doi.org/10.1145/3377325.3377519>
- [6] Grigoris Antoniou, Katie Atkinson, George Baryannis, Sotiris Batsakis, Luigi Di Caro, Guido Governatori, Livio Robaldo, Giovanna Siragusa, and Ilias Tachmazidis. 2022. Explainable Reasoning with Legal Big Data: A Layered Framework. *Journal of Applied Logic* 9 (2022), 1155–170.
- [7] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschanen, Emanuele Bugliarello, et al. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726* (2024).
- [8] Shannon Bond. 2024. *How deepfakes and AI memes affected global elections in 2024*. <https://www.npr.org/2024/12/21/nx-s1-5220301/deepfakes-memes-artificial-intelligence-elections> Accessed: 2025-04-12.
- [9] Bartosz Brożek, Michał Furman, Marek Jakubiec, and Bartłomiej Kucharzyk. 2024. The black box problem revisited. Real and imaginary challenges for automated legal decision making. *Artificial Intelligence and Law* 32, 2 (2024), 427–440. <https://doi.org/10.1007/s10506-023-09356-9>
- [10] Case IQ. 2024. *What is Deepfake Identity Theft?* <https://www.caseiq.com/resources/what-is-deepfake-identity-theft/> Accessed: 2025-04-12.
- [11] Ilias Chalkidis and Dimitrios Kampas. 2019. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artif. Intell. Law* 27, 2 (2019), 171–198. <https://doi.org/10.1007/s10506-018-9238-9>
- [12] Heather Chen and Kathleen Magrino. 2024. Finance worker pays out \$25 million after video call with deepfake CFO. CNN (2024). <https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html> Accessed: 2025-04-10.
- [13] Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, et al. 2024. When large language models meet personalization: Perspectives of challenges and opportunities. *World Wide Web* 27, 4 (2024), 42.
- [14] Yuwei Chen. 2025. Towards Human Explainable Digital Forensics: Generating Human Interpretable Evidence for Semantic Understanding in Manipulated Images and Text. (2025).
- [15] Beomsang Cho, Binh M Le, Jiwon Kim, Simon Woo, Shahroz Tariq, Alsharif Abuadbba, and Kristen Moore. 2023. Towards understanding of deepfake videos in the wild. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 4530–4537.
- [16] François Fleuret. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1251–1258.
- [17] Ben Colman. 2025. *2,100% Surge in AI Fraud as Deepfake Regulation Lags.* <https://www.realitydefender.com/blog/2-100-surge-in-ai-fraud-as-deepfake-regulation-lags> Accessed: 2025-04-12.
- [18] Hoang Cuong Nguyen, Shahroz Tariq, Mohan Baruwal Chhetri, and Bao Quoc Vo. 2025. Towards effective identification of attack techniques in cyber threat intelligence reports using large language models. In *Companion Proceedings of the ACM on Web Conference 2025*. 942–946.
- [19] Shreyansh Dani and contributors. 2024. VisionGPT2: A Vision-to-Language Model. <https://github.com/shreydan/VisionGPT2>. Accessed: 2024-12-23.
- [20] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. 2023. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3994–4004.
- [21] Joel Frank, Thorsten Eisenhofer, Lea Schönher, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. 2020. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*. PMLR, 3247–3258.
- [22] Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. 2023. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems* 36 (2023), 5539–5568.
- [23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.
- [24] Brian Hosler, Davide Salvii, Anthony Murray, Fabio Antonacci, Paolo Bestagini, Stefano Tubaro, and Matthew C Stamm. 2021. Do deepfakes feel emotions? A semantic approach to detecting deepfakes via emotional inconsistencies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1013–1022.
- [25] Mousa Jafar, Mohammad Ababneh, Mohammed Al-Zoube, and Ammar Elhassan. 2020. *Forensics and Analysis of Deepfake Videos*. 053–058 pages. <https://doi.org/10.1109/ICICS4946.2020.930493>
- [26] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.
- [27] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon S Woo. 2021. Evaluation of an Audio-Video Multimodal Deepfake Dataset using Unimodal and Multimodal Detectors. In *Proceedings of the 1st Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection*. 7–15.
- [28] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. 2021. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- [29] Jeongho Kim, Shahroz Tariq, and Simon S Woo. 2022. PTD: Privacy-Preserving Human Face Processing Framework using Tensor Decomposition. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. 1296–1303. <https://doi.org/10.1145/3477314.3507036>
- [30] Minha Kim, Shahroz Tariq, and Simon S Woo. 2021. Cored: Generalizing fake media detection with continual representation using distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 337–346.
- [31] Minha Kim, Shahroz Tariq, and Simon S Woo. 2021. FReTAL: Generalizing Deepfake Detection using Knowledge Distillation and Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1001–1012.
- [32] Vejay Lalla, Adine Mitrani, and Zach Harned. 2022. Artificial Intelligence: Deepfakes in the Entertainment Industry. *WIPO Magazine* (2022). <https://www.wipo.int/web/wipo-magazine/articles/artificial-intelligence-deepfakes-in-the-entertainment-industry-42620> Accessed: 2025-04-10.
- [33] Binh Le, Shahroz Tariq, Alsharif Abuadbba, Kristen Moore, and Simon Woo. 2023. Why Do Facial Deepfake Detectors Fail?. In *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes* (Melbourne, VIC, Australia) (WDC ’23). Association for Computing Machinery, New York, NY, USA, 24–28. <https://doi.org/10.1145/3595353.3595882>
- [34] Binh M. Le, Jiwon Kim, Simon S. Woo, Kristen Moore, Alsharif Abuadbba, and Shahroz Tariq. 2025. SoK: Systematization and Benchmarking of Deepfake Detectors in a Unified Framework. *arXiv:2401.04364 [cs.CV]* <https://arxiv.org/abs/2401.04364>
- [35] Sangyup Lee, Shahroz Tariq, Junyaup Kim, and Simon S Woo. 2021. TAR: Generalized Forensic Framework to Detect Deepfakes Using Weakly Supervised Learning. In *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer, 351–366.
- [36] Sangyup Lee, Shahroz Tariq, Youjin Shin, and Simon S Woo. 2021. Detecting handcrafted facial image manipulations and GAN-generated facial images using Shallow-FakeFaceNet. *Applied Soft Computing* 105 (2021), 107256.
- [37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- [38] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.
- [39] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. [n. d.]. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5001–5010.
- [40] Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. 2022. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems* 64, 12 (2022), 3197–3234.
- [41] Y Li. 2018. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656* (2018).

- [42] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3207–3216.
- [43] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*. Springer, 740–755.
- [44] Marco Lippi, Przemysław Palka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. 2019. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law* 27 (2019), 117–139.
- [45] Jiawei Liu, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. 2024. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. *arXiv preprint arXiv:2410.10238* (2024).
- [46] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August 15, 2018* (2018), 11.
- [47] Scott Lundberg. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).
- [48] Arpan Mahara and Napthalit Risse. 2025. Methods and Trends in Detecting Generated Images: A Comprehensive Review. *arXiv preprint arXiv:2502.15176* (2025).
- [49] Manoj Kumar Mahto, Durgesh Srivastava, Ranjit Kumar, Basant Sah, Hare Ram Singh, and Sunil Kr. Maakar. 2025. Personalized User Interaction in Web Applications using Adaptive LLM Model. In *2025 International Conference on Pervasive Computational Technologies (ICPCT)*. 962–966. <https://doi.org/10.1109/ICPCT4145.2025.10940477>
- [50] Asad Malik, Minoru Kuribayashi, Sani M. Abdullahi, and Ahmad Neyaz Khan. 2022. DeepFake Detection for Human Face Images and Videos: A Survey. *IEEE Access* 10 (2022), 18757–18775. <https://doi.org/10.1109/ACCESS.2022.3151186>
- [51] Yisroel Mirsky and Wenke Lee. 2021. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)* 54, 1 (2021), 1–41.
- [52] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*. 2823–2832.
- [53] Fania Mokhayeri, Eric Granger, and Guillaume-Alexandre Bilodeau. 2015. Synthetic face generation under various operational conditions in video surveillance. In *2015 IEEE International Conference on Image Processing (ICIP)*. 4052–4056. <https://doi.org/10.1109/ICIP.2015.7351567>
- [54] Yunsheng Ni, Depu Meng, Changqian Yu, Chengbin Quan, Dongchun Ren, and Youjian Zhao. [n. d.]. Core: Consistent representation learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12–21.
- [55] Cécile Paris and Andrew Reeson. 2024. What's the Secret to Making Sure AI Does Not Steal Your Job? Work with It, Not Against It. In *The Conversation on Work*. Ian O. Williamson (Ed.). Johns Hopkins University Press, Baltimore, 177–181. First published in The Conversation on November 30th, 2021.
- [56] Yogesh Patel, Sudeep Tanwar, Rajesh Gupta, Pronaya Bhattacharya, Innocent Ewean Davidson, Royi Nyameko, Srinivas Aluvula, and Vrince Vimal. 2023. Deepfake Generation and Detection: Case Study and Challenges. *IEEE Access* 11 (2023), 143296–143323. <https://doi.org/10.1109/ACCESS.2023.3342107>
- [57] D Philmee. 2023. Practice Innovations: Seeing is no longer believing—The rise of deepfakes. *Thomson Reuters* 18 (2023).
- [58] Ben Pinhasov, Raz Lapid, Rony Ohayon, Moshe Sipper, and Yehudit Aperstein. 2024. Xai-based detection of adversarial attacks on deepfake detectors. *arXiv preprint arXiv:2403.02955* (2024).
- [59] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*. Springer, 86–103.
- [60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [61] Victoria Reed. 2024. *The Bright Side of Deepfakes: Edu, Films & Digital Legacies*. <https://aicompetence.org/the-bright-side-of-deepfakes-education-films/> Accessed: 2025-04-12.
- [62] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. , 1135–1144 pages. <https://doi.org/10.1145/2939672.2939778>
- [63] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1–11.
- [64] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
- [65] Samuel Henrique Silva, Mazal Bethany, Alexis Megan Votto, Ian Henry Scarff, Nicole Beebe, and Peyman Najafirad. 2022. Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Science International: Synergy* 4 (2022), 100217.
- [66] Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. 2025. Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 952–966. <https://doi.org/10.1145/3708359.3712091>
- [67] Razaib Tariq, Minji Heo, Simon S. Woo, and Shahroz Tariq. 2024. Beyond the Screen: Evaluating Deepfake Detectors under Moiré Pattern Effects. In *Workshop on Media Forensics (CVPR)*. Seattle.
- [68] Razaib Tariq, Simon S. Woo, and Shahroz Tariq. 2024. Exploring the Impact of Moiré Pattern on Deepfake Detectors. In *Special Session Recent Advances in Multimedia Forensics in the AI-Generative Era (ICIP)*. Abu Dhabi, UAE.
- [69] Shahroz Tariq, Alsharif Abuabdulla, and Kristen Moore. 2023. Deepfake in the Metaverse: Security Implications for Virtual Gaming, Meetings, and Offices. In *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes* (Melbourne, VIC, Australia) (WDC '23). Association for Computing Machinery, New York, NY, USA, 16–19. <https://doi.org/10.1145/3595353.3595880>
- [70] Shahroz Tariq, Mohan Baruwal Chhetri, Surya Nepal, and Cecile Paris. 2025. A2C: A modular multi-stage collaborative decision framework for human–AI teams. *Expert Systems with Applications* 282 (2025), 127318.
- [71] Shahroz Tariq, Sowon Jeon, and Simon Woo. 2021. Am I a Real or Fake Celebrity? Measuring Commercial Face Recognition Web APIs under Deepfake Impersonation Attack. *arXiv preprint arXiv:2103.00847* (2021).
- [72] Shahroz Tariq, Sowon Jeon, and Simon S. Woo. 2023. Evaluating Trustworthiness and Racial Bias in Face Recognition APIs Using Deepfakes. *Computer* 56, 5 (2023), 51–61.
- [73] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S. Woo. 2018. Detecting both machine and human created fake face images in the wild. In *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*. ACM, 81–87.
- [74] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S. Woo. 2019. GAN is a friend or foe?: a framework to detect various fake face images. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*. ACM, 1296–1303.
- [75] Shahroz Tariq, Sangyup Lee, and Simon Woo. 2021. One detector to rule them all: Towards a general deepfake attack detection framework. In *Proceedings of the web conference 2021*.
- [76] Shahroz Tariq, Sangyup Lee, and Simon S. Woo. 2020. A Convolutional LSTM based Residual Network for Deepfake Video Detection. *arXiv preprint arXiv:2009.07480* (2020).
- [77] Shahroz Tariq, David Nguyen, MAP Chamikara, Tingmin Wu, Alsharif Abuabdulla, and Kristen Moore. 2025. LLMs Are Not Yet Ready for Deepfake Image Detection. *arXiv preprint arXiv:2506.10474* (2025).
- [78] Shahroz Tariq, Priyanka Singh, and Simon S. Woo. 2025. DF-P2E Framework Demo: Interactive Deepfake Detection for Non-Expert Users. <https://doi.org/10.5281/zenodo.15198666>
- [79] Shahroz Tariq, Ronal Singh, Mohan Baruwal Chhetri, Surya Nepal, and Cecile Paris. 2025. Bridging expertise gaps: The role of llms in human-ai collaboration for cybersecurity. *arXiv preprint arXiv:2505.03179* (2025).
- [80] Mulin Tian, Mahyar Khayatkhoei, Joe Mathai, and Wael AbdAlmageed. 2023. Unsupervised multimodal deepfake detection using intra-and cross-modal inconsistencies. *arXiv preprint arXiv:2311.17088* (2023).
- [81] Hugo Touvron, Thibaut Lavrille, Gautier Izard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, and Faisal Azhar. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [82] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100* (2022).
- [83] Luozhou Wang, Shuai Yang, Shu Liu, and Ying-cong Chen. 2023. Not All Steps are Created Equal: Selective Diffusion Distillation for Image Manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7472–7481.
- [84] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*. PMLR, 23318–23340.
- [85] Saima Waseem, Syed Abdul Rahman Syed Abu Bakar, Bilal Ashfaq Ahmed, Zaid Omar, Taiseer Abdalla Elfadil Eisa, and Mhassen Elnoor Elneel Dalam. 2023. DeepFake on Face and Expression Swap: A Review. *IEEE Access* 11 (2023), 117865–117906. <https://doi.org/10.1109/ACCESS.2023.3324403>
- [86] Jun Wei, Shuhui Wang, and Qingming Huang. 2020. F³Net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 12321–12328.

- [87] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. 2022. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 657–666.
- [88] Wenju Xu, Shawn Keshmiri, and Guanghui Wang. 2019. Adversarially Approximated Autoencoder for Image Generation and Manipulation. *IEEE Transactions on Multimedia* 21, 9 (2019), 2387–2396. <https://doi.org/10.1109/TMM.2019.2898777>
- [89] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Li Yuan, Chengjie Wang, Shouhong Ding, et al. 2024. DF40: Toward Next-Generation Deepfake Detection. *arXiv preprint arXiv:2406.13495* (2024).
- [90] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 8261–8265.
- [91] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing* 4, 2 (2024), 100211. <https://doi.org/10.1016/j.hcc.2024.100211>