Trustworthy Image Authentication using Forensic Knowledge Graphs

Tai D. Nguyen[1] and Matthew C. Stamm[1]

Drexel University, Philadelphia, PA 19104, USA

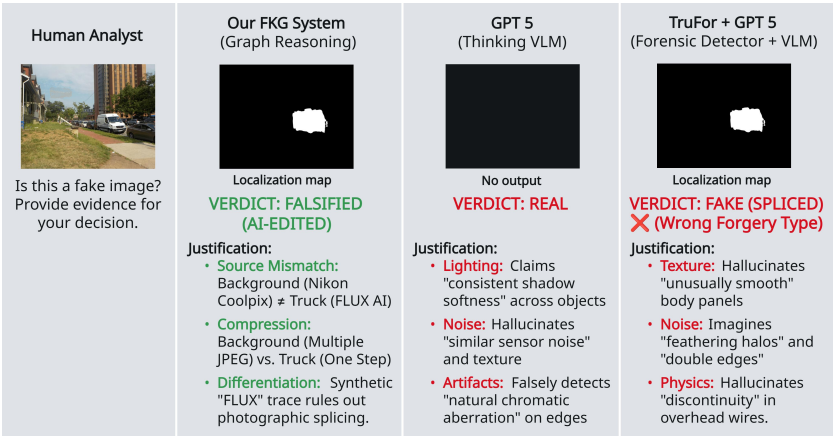

Fig. 1: Unlike existing systems that miss forgeries or hallucinate evidence, our FKG framework delivers trustworthy authentication grounded in real forensic traces.

Abstract. Advances in generative AI have made image falsification highly realistic, demanding trustworthy authentication systems. Existing forensic detectors can target certain forgery types but lack interpretability, while vision-language models (VLMs) provide explanations but cannot exploit forensic traces for reliable detection. We propose Forensic Knowledge Graphs (FKGs), a unified framework that integrates forensic evidence extraction, structured reasoning, and human-interpretable explanation. Our FKG structure encodes forensic traces along with their causal dependencies and links to scene content. To generate accurate FKGs, we introduce a novel forensic authentication network and an Iterative Context Refinement strategy that guides VLMs to produce faithful, grounded explanations. We also present FKG-50K, a dataset of 50,000 realistic forgeries with ground-truth FKGs. Experiments demonstrate that FKG outperforms both forensic detectors and VLMs in detection, forgery identification and localization, and forensic justification.

Keywords: Image Forensics · Knowledge Graphs · Trustworthy AI

1 Introduction

Recent advances in generative AI [30, 67] and modern AI image editors [8, 12, 27, 59, 65] have dramatically expanded the ways images can be falsified. Hyperrealistic forgeries can now be produced through fully AI-generated imagery [40, 63, 66], content splicing, localized editing, and AI-based modifications [79]. As

these capabilities improve, distinguishing authentic images from falsified ones has become increasingly challenging, even for expert analysts [32, 55, 56].

To protect against these threats, there is a significant need for media authentication systems [79]. However, for these systems to be effective in practice, users must be able to trust them [44]. Specifically, users require interpretable and explainable outputs to establish trust in automated decision-making systems [7, 23, 35], no matter how accurate they are. Therefore, a trustworthy authentication system must: (1) provide **accurate, general-purpose detection** regardless of forgery type, (2) produce **human-interpretable explanations** describing what was falsified and how, and (3) provide **justification of forensic reasoning** that grounds every claim in verifiable evidence.

Current authentication systems only partially satisfy these requirements. While numerous specialized systems have been developed, such as synthetic image detectors [74, 81], manipulation classifiers [10, 36], and splicing localizers [18, 34], each method only targets a narrow family of forgeries and fails to generalize beyond it. For example, a splicing detector may achieve high accuracy on spliced images but miss AI-generated edits entirely [1, 50], while a synthetic image detector may flag fully generated content but overlook traditional manipulations [19, 50]. At the moment, no single system can accurately detect a broad spectrum of forgery types [68, 79]. Furthermore, their outputs are often confined to binary "real/fake" labels or heatmaps of suspicious areas [18, 34], with no additional explanation of what content is falsified, how it was manipulated, or why the system believes a particular region is inauthentic [47, 83].

At the opposite extreme, vision-language models (VLMs) [28, 46, 58] can produce natural language justifications, but utilize visual or semantic inconsistencies [39, 80] such as unnatural lighting, cues rapidly disappearing as AI generators produce increasingly realistic outputs [84, 88]. More critically, VLMs cannot leverage the invisible statistical traces (*i.e.*, forensic microstructures such as sensor noise, compression artifacts, and demosaicing residuals) that reliable forensic techniques rely on [20, 70]. Recent benchmarks confirm that even the best VLMs barely exceed chance-level accuracy on forgery detection tasks [39, 75, 80], and in real-world incidents, chatbots have confidently declared AI-generated images authentic [24, 26], eroding public trust in digital media and AI verification [25].

To address these challenges, we introduce Forensic Knowledge Graphs (FKGs), a unified framework for trustworthy image authentication that integrates forensic evidence, structured reasoning, and interpretable explanation within a single system. FKGs capture forensic analyses as a structured graph representation, linking image regions with their sources, manipulation histories, and relevant scene content. This representation allows the system to analyze multiple forgery types simultaneously, preserve dependencies among forensic traces, link evidence to scene regions, and generate grounded natural-language explanations. We perform comprehensive experimental analysis of our framework using multiple datasets, demonstrating that FKG outperforms both existing forensic detectors and VLMs in detection, forgery identification and localization, and forensic justification. Our contributions are summarized as follows:

1. We introduce *Forensic Knowledge Graphs*, a novel information structure that encodes forensic evidence, causal dependencies among analyses, and links findings to semantic scene content.
2. We propose a *unified forensic authentication network* composed of a novel self-supervised trace extraction backbone, a novel hybrid region proposal network, and a transformer-based reasoner that can detect multiple different forgery types and construct accurate FKGs from input images.
3. We propose *Iterative Context Refinement (ICR)*, a novel prompting strategy that enables vision-language models to generate faithful, forensic-evidence-grounded summaries and justifications from FKGs.
4. We release *FKG-50K*, a dataset of 50,000 images with ground-truth FKGs covering multiple forgery types, and demonstrate state-of-the-art performance across all three dimensions of trustworthy authentication.

2 Background & Related Work

Forensic Microstructures. These are subtle, content-independent pixel correlations imprinted by an image's acquisition and processing history [70, 72]. They arise from sensor noise [49], demosaicing [87], and compression [61, 71], collectively forming a statistical "fingerprint" unique to each imaging pipeline [51]. Unlike semantic cues, these traces persist across diverse content, enabling forensic systems to infer an image's origin [9, 10], detect inconsistencies among regions [18, 20], and distinguish camera artifacts from AI-generated ones [19, 81].

VLMs for Image Authentication. As AI-generated imagery proliferates online [25], users and journalists increasingly turn to VLMs such as ChatGPT and Gemini to verify image authenticity [17,24]. However, as discussed in Sec. 1, even the best VLMs barely exceed chance-level accuracy on forensic tasks [39, 75, 80], and fact-checking investigations confirm that VLMs and detection tools fail under simple perturbations like cropping or rescaling [2, 24]. Even forensic VLMs like FakeShield [83], ForgeryGPT [47], SIDA [38], and recent reasoning-oriented detectors [4, 14, 73] rely on visible cues rather than forensic microstructures, and cover only narrow forgery families (ForgeryGPT handles only faces).

Forensic Systems. Modern forensic systems exploit these microstructures but vary in scope and design [79]. Traditional detectors like CAT-Net [43], HiFi-IFDL [36], and TruFor [34] focus on specific manipulation families such as splicing and compression inconsistencies, while newer approaches target AI-generated content by modeling generator-specific artifacts (*e.g.*, NPR [74], DE-FAKE [69], UFD [57], FSD [54]). However, these systems operate as binary detectors that cannot reason about causal relationships between multiple forensic cues [1, 79] struggling with forgeries that blend real and synthetic content [50].

3 Forensic Knowledge Graphs

We introduce Forensic Knowledge Graphs (FKGs) - a hierarchical structure whose nodes represent forensic entities and edges encode labeled relations capturing how findings are interconnected and why they support a particular authenticity conclusion. The FKG's structure and semantics are defined by an

ontology that specifies entity classes and their permissible relationships [33]. An overview follows with formal definitions in Appendix A.

Entity Classes and Relations. The ontology defines six core entity classes and their typed relations: (1) **Image** - the root node representing the analyzed image; (2) **Region** - a pixel subset sharing common forensic traces, serving as the central entity from which all relations originate; (3) **Source** - the origin of pixels, with subclasses for *Camera Model* and *AI Generator*; (4) **Post-Processing** - operations applied after generation, *e.g.* resampling, blur, noise; (5) **Compression** - compression artifacts, with subclasses tracking whether a region is uncompressed, its last compression parameters, and whether recompression occurred; and (6) **Content** - semantic scene elements characterized by objectedness and object description.

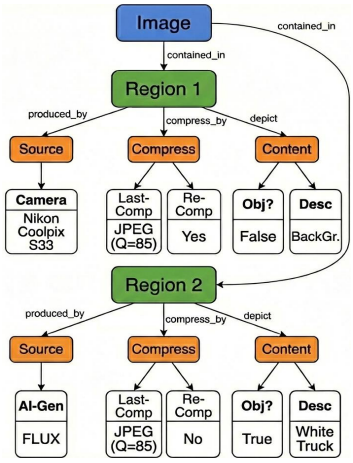

Fig. 2: FKG for the AI-edited image in Fig. 1.

Relations encode dependencies among entity classes as ontology object properties. Each Region is linked to its parent Image via `contained_in`, to its origin via `produced_by`, to post-processing and compression nodes via `modified_by` and `compress_by`, and to semantic content via `depict`. Additional datatype properties (Appendix A) associate entities with descriptive literals like probabilities, model IDs, and post-processing parameters. These typed edges are instantiated deterministically from our system's per-region predictions, not inferred by a VLM, so every relation is grounded in an explicit forensic prediction.

Illustrative FKG Example. Fig. 2 shows the FKG for an AI-edited image where a truck was added to a real photo. The graph contains two Region nodes: one linked to Nikon Coolpix S33 and another to FLUX generator. Compression evidence reinforces this split: the camera region shows multiple recompressions while the FLUX region shows only one, indicating post-hoc content insertion.

4 FKG Generation System

Overview. Our framework (Fig. 3) unifies forensic evidence extraction, structured reasoning, and interpretable explanation in a single pipeline. Given an input image, the FKG Generation System (Fig. 4) divides it into patches and extracts forensic fingerprints via a Forensic Backbone Network. A Forensic Region Proposal Network (FRPN) then clusters patches with similar fingerprints into forensic regions. Finally, Task Expert Networks and a Transformer Reasoning Module infer each region's forensic attributes and map them to the FKG ontology. A VLM then translates the resulting FKG into a human-interpretable forensic explanation, guided by our Iterative Context Refinement strategy (Sec. 5). Full architectural and training details are in Appendix B–D.

Forensic Backbone Network. Generating an FKG first requires a backbone network that captures forensic microstructures, statistical fingerprints left by an image's source and processing history [20, 70]. These *general-purpose* embeddings are the foundation of our system: the FRPN relies on them to partition the image into forensic regions, and the task experts rely on them to infer forensic attributes. Prior methods learn these through task-specific supervision (*e.g.*, camera model classification) [9, 10, 18, 36, 43], tying the backbone to specific sources and limiting generalization [1, 50, 79].

Rather than learning from task-specific labels, we exploit a fundamental yet surprisingly simple property of imaging pipelines: *local patches from the same image share a common forensic fingerprint, while patches from different images do not.* This observation alone turns every unlabeled image into a free training sample, enabling the backbone to learn general forensic fingerprints at scale without forgery labels or source annotations. To prevent the network from learning content-dependent shortcuts (*e.g.*, mapping semantically similar patches to the same image), we first process each patch using a learned high-pass filter [9, 10] to suppress image content. We then enforce self-supervision through two complementary objectives: a pairwise similarity term enforcing same-image alignment and cross-image separation, and a contrastive term sharpening per-patch discrimination. Given a set of patches $\mathcal{X}$, the backbone is trained with:

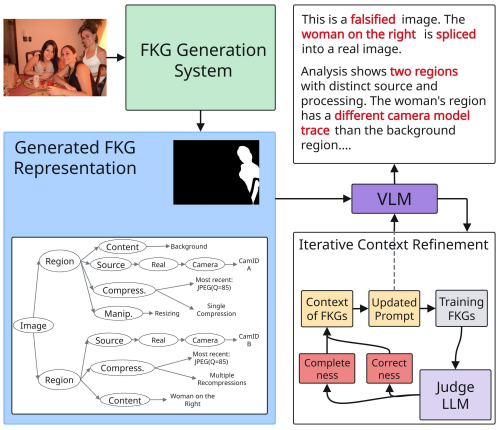

Fig. 3: Overview of our full FKG framework.

$$\mathcal{L}_B = \frac{1}{|\mathcal{X}|^2} \sum_{i,j} w_{ij} (Y_{ij} - z_i^\top z_j)^2 - \lambda \sum_i \log \frac{e^{\cos(z_i, z_{i^+})/\tau}}{\sum_j e^{\cos(z_i, z_j)/\tau}}, \tag{1}$$

where z_i are ℓ_2-normalized patch embeddings, $Y_{ij} \in \{0, 1\}$ is a same-image indicator, z_{i^+} is the positive pair of z_i, w_{ij} balance positive and negative pairs, and τ is a temperature. These embeddings encode each patch's forensic identity and serve as input to the FRPN.

Forensic Region Proposal Network. Given the backbone's forensic embeddings, the next step is partitioning the image into regions that share a common forensic trace. Existing visual segmentation and region proposal methods [37, 42] cannot accomplish this because they operate on visual features, whereas forensic regions are defined by *invisible* statistical properties: a spliced region may appear visually seamless yet carry an entirely different forensic fingerprint.

To address this, we introduce a Forensic Region Proposal Network (FRPN) based on the following observation: *partitioning an image into forensically homogeneous regions requires comparing every patch against every other — the*

all-pairs operation that transformer self-attention naturally computes. However, global self-attention suffers from *attention dilution* [77]: accumulated contributions from distant, unrelated patches cause forensically distinct regions to appear connected, leading to false alarms and misclassified clusters. To mitigate this, we propose a novel *Hybrid Graph Attention Transformer* that takes the backbone embeddings $Z = [z_1, \ldots, z_{|\mathcal{P}|}]$ as input tokens and alternates global self-attention with local graph attention. Local graph attention suppresses false alarms by restricting aggregation to each token's k-nearest neighbors using an additive scoring mechanism [78]. The attention weights under each mode are:

$$\alpha_{ij}^{\text{global}} = \frac{\exp(\mathbf{w}_{qi}^\top \mathbf{w}_{kj}/\sqrt{d})}{\sum_{m=1}^{|\mathcal{P}|} \exp(\mathbf{w}_{qi}^\top \mathbf{w}_{km}/\sqrt{d})}, \quad \alpha_{ij}^{\text{local}} = \frac{\exp\big(\sigma(\mathbf{a}_s^\top \mathbf{w}_{gi} + \mathbf{a}_d^\top \mathbf{w}_{gj})\big)}{\sum_{m\in\mathcal{N}(i)} \exp\big(\sigma(\mathbf{a}_s^\top \mathbf{w}_{gi} + \mathbf{a}_d^\top \mathbf{w}_{gm})\big)}, \tag{2}$$

where $\mathbf{w}_{ni} = W_n z_i \in \mathbb{R}^d$ are learned projections of token i, σ is the LeakyReLU activation, $\mathbf{a}_s, \mathbf{a}_d$ are learned score vectors, and $\mathcal{N}(i)$ are the dynamic k-nearest neighbors of i by cosine similarity. Stacking L alternating layers yields: $\tilde{Z} = \text{Attn}_g^{(L)} \circ \text{Attn}_\ell^{(L)} \circ \cdots \circ \text{Attn}_g^{(1)} \circ \text{Attn}_\ell^{(1)}(Z)$, where Attn_g and Attn_ℓ denote global self-attention and local graph attention layers. The output $\tilde{Z} = [\tilde{z}_1, \ldots, \tilde{z}_{|\mathcal{P}|}]$ are refined embeddings from which a classification head predicts a cluster assignment $\hat{c}_i$ for each patch.

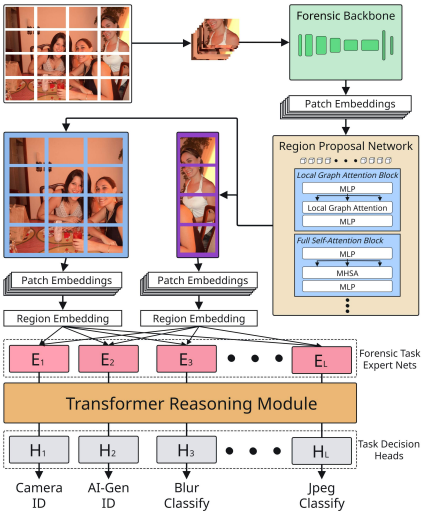

Fig. 4: Overview of our system which generates a FKG for an input image.

To train cluster assignments against ground-truth forensic regions (*e.g.*, from known tampering masks), we must account for the fact that predicted cluster IDs are arbitrary and do not correspond to ground-truth region labels. We use Hungarian matching [15] to find the optimal assignment, and define a loss by combining it with a pairwise similarity term that preserves embedding structure:

$$\mathcal{L}_F = \frac{1}{|\mathcal{P}|^2} \sum_{i,j} w_{ij}(Y_{ij} - \tilde{z}_i^\top \tilde{z}_j)^2 + \alpha \sum_{i=1}^{|\mathcal{P}|} \ell_{\text{CE}}\big(y_i,\ \hat{\sigma}(\hat{c}_i)\big), \tag{3}$$

where $\tilde{z}_i$ are ℓ_2-normalized, $Y_{ij} = \mathbf{1}[y_i = y_j]$ is a same-region indicator, ℓ_{CE} is the cross-entropy loss, $\hat{\sigma}$ is the optimal Hungarian matching from predicted to ground-truth labels [15], and α balances the two terms. Patches sharing the same predicted cluster form a forensic region, serving as Region nodes in the FKG.

Forensic Task and Decision Making. Given the forensic regions produced by the FRPN, this stage determines *what happened* to each region and assembles the final FKG. Each region's refined patch embeddings $\{\tilde{z}_i\}$ are average-pooled into a region-level vector ψ_k, then processed by Forensic Task Expert Networks - MLPs specializing in source identification, post-processing classification, and compression analysis - each producing a task-specific embedding $\xi_k^{(t)}$.

A Transformer Reasoning Module refines these by modeling cross-task dependencies (*e.g.*, compression artifacts informing source identification) to produce updated features $\tilde{\xi}_k^{(t)}$, from which task-specific heads predict binary, categorical, or continuous outputs under: $\mathcal{L}_T = \sum_{t \in \mathcal{T}} \lambda_t \mathcal{L}_t$, where $\mathcal{T}$ is the set of forensic tasks, λ_t are task-specific weights, and $\mathcal{L}_t$ is cross-entropy or mean-squared error between predictions from $\tilde{\xi}_k^{(t)}$ and ground-truth labels.

These predictions are mapped to the FKG ontology: each region becomes a Region node whose predicted attributes instantiate entity nodes and edges via `produced_by`, `modified_by`, and `compress_by` relations. When different regions link to different sources, the graph directly reveals manipulation. The backbone, FRPN, and task networks are trained sequentially (details in Appendix D).

5 Interpreting and Reasoning over FKGs

Given an FKG, we use a vision-language model (VLM) to produce forensic decisions from its region-level evidence, including image authenticity, manipulation localization, and whether content is spliced, AI-edited, or fully synthetic, along with justifications grounding each decision in FKG evidence.

This requires addressing two challenges. First, the FKG combines symbolic attributes with references to spatial visual content, making it inherently multimodal, and standard VLMs cannot consume graph structures directly. Second, even with a serialized FKG, current VLMs lack forensic domain knowledge and tend to fabricate or omit evidence, yielding untrustworthy justifications.

We serialize the FKG F into subject-predicate-object triplets $F = \{(s_i, p_i, o_i)\}$. However, triplets alone cannot convey what each region depicts. We therefore use Set-of-Mark (SoM) prompting [85], overlaying each Region with a numbered mark on the image and referencing it symbolically in the triplets (Fig. 5, top). This visual grounding enables the VLM to recognize the scene content (*e.g.*, a spliced object or background) and link it to forensic evidence. These, along with task instructions and a response schema (Appendix F), form the input prompt to the VLM.

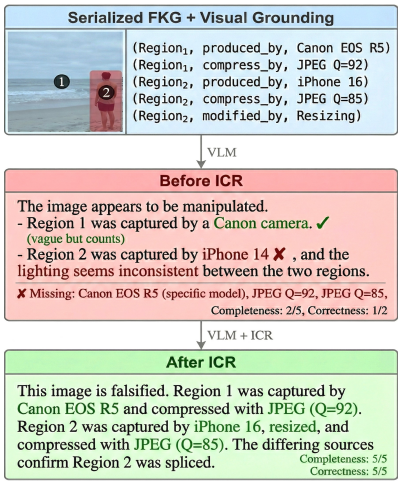

Fig. 5: ICR improves forensic justifications by refining VLM's context.

Iterative Context Refinement. A naive approach to the second challenge is in-context learning [13]. However, manual curation of good FKG-report pairs is infeasible given FKG diversity, and existing prompt optimization methods [41, 45, 64, 86] optimize for producing correct answers (*e.g.*, to a math problem), not for faithfulness to structured evidence. We therefore propose Iterative Context Refinement (ICR), a novel optimization strategy that automatically curates in-context examples to maximize report faithfulness and completeness. To measure these quantities, we treat each

triplet $m_j \in F$ as an atomic fact and define two indicators: $\tau(m_j, R) \in \{0, 1\}$ for whether m_j is mentioned in report R, and $Q(m_j, R) \in \{0, 1\}$ for whether it is correctly stated. Then, we define completeness and correctness, as:

$$M_{\text{comp}} = \frac{1}{|F|} \sum_j \tau(m_j, R), \quad M_{\text{corr}} = \frac{1}{\sum_j \tau(m_j, R)} \sum_j Q(m_j, R)\tau(m_j, R). \tag{4}$$

The key idea behind ICR is to identify and correct the VLM's failure modes. Starting from $\mathcal{C} = \emptyset$ (Fig. 3), a judge LM [89] evaluates M_{comp} and M_{corr} for each FKG in a training set of N authentic and manipulated images. The FKGs where the VLM most severely fabricates or omits evidence are paired with their errors as corrective demonstrations and appended to $\mathcal{C}$, forming an updated context set $\mathcal{C}'$ that is prepended to the VLM's prompt in the next round. By focusing on the VLM's blind spots, ICR progressively drives faithfulness toward convergence:

$$\frac{1}{N} \sum_i \left(|M^{(t)}_{\text{corr},i} - M^{(t-1)}_{\text{corr},i}| + |M^{(t)}_{\text{comp},i} - M^{(t-1)}_{\text{comp},i}| \right) < \epsilon. \tag{5}$$

As shown in Fig. 5 (bottom), this ensures every forensic claim is backed by verifiable graph evidence and every piece of evidence is accounted for.

6 Datasets

Evaluating an image authentication system requires ground-truth data whose provenance is ***fully known, accurately labeled, and completely captured***. Existing forensic datasets provide at most binary real/fake labels or tampering masks, which is insufficient for evaluating systems that reason over structured forensic evidence such as source identity, compression history, and modification lineage. To address this, we curate **FKG-50K**, a dataset of 50,000 images with complete provenance and manipulation histories, built from a corpus of pristine, camera-original photographs sourced from Unsplash [76], Pexels [62], and crowdsourcing. We filtered for images with rich EXIF metadata (camera model, processing parameters, color pipeline), as these define the source-level forensic fingerprint the FKG ontology captures. Details are provided in Appendix E.

From these originals, we generated *realistic* manipulated counterparts spanning splicing, local retouching, and AI-based editing (FLUX-Kontex-1 [8], SDXL-Inpaint [63]), recording each manipulation as a structured operation encoded into its corresponding FKG. We also generated *photo-realistic* fully synthetic counterparts from originals using SDXL [63], FLUX [11], Midjourney v6 [53], & GPT-Image-1 [59]. The dataset includes 40K training and 10K evaluation samples. We also benchmark on out-of-distribution public datasets such as DSO-1 [16], CASIAv2 [22], GenImage [90], Synthbuster [6].

7 Experiments

We evaluate our FKG framework & others on three dimensions of trustworthy image authentication: (1) detecting falsified images, (2) identifying how & where manipulations occur, (3) providing interpretable, evidence-based justifications.

Table 1: Fake & manipulated image detection on FKG-50K and other datasets. Each cell shows ACC / AUC. [I] = Instruct, [T] = Thinking.

Method	FKG-50K					Out of Distribution Datasets			
						Splice & Edit		Fully AI-Gen	
	AI-Edit	Full AI	Splicing	Trad-Edit	Overall	DSO-1	CASIAv2	GenImage	Synthbuster
GPT-5	0.59 / 0.60	0.68 / 0.82	0.81 / 0.84	0.55 / 0.61	0.66 / 0.72	0.59 / 0.61	0.74 / 0.75	**0.93** / 0.88	0.90 / 0.86
Sonnet 4	0.55 / 0.52	0.72 / 0.75	0.54 / 0.53	0.51 / 0.51	0.58 / 0.58	0.50 / 0.53	0.56 / 0.62	0.85 / 0.66	0.91 / 0.78
Gemini 2.0	0.53 / 0.51	0.64 / 0.62	0.71 / 0.72	0.55 / 0.59	0.60 / 0.61	0.52 / 0.53	0.66 / 0.65	0.90 / 0.75	0.91 / 0.78
Qwen3 VL [I]	0.53 / 0.53	0.72 / 0.71	0.63 / 0.63	0.54 / 0.54	0.60 / 0.60	0.50 / 0.51	0.59 / 0.59	0.89 / 0.73	0.92 / 0.81
Qwen3 VL [T]	0.57 / 0.57	0.69 / 0.80	0.68 / 0.72	0.54 / 0.53	0.62 / 0.66	0.52 / 0.53	0.57 / 0.61	0.61 / 0.84	0.60 / 0.84
Gemma 3	0.56 / 0.58	0.69 / 0.76	0.65 / 0.68	0.50 / 0.52	0.60 / 0.64	0.57 / 0.57	0.52 / 0.55	0.69 / 0.75	0.69 / 0.79
Llama 4 S.	0.52 / 0.52	0.64 / 0.65	0.52 / 0.50	0.50 / 0.52	0.54 / 0.55	0.48 / 0.51	0.56 / 0.57	0.73 / 0.68	0.71 / 0.58
Nemotron N.	0.52 / 0.52	0.48 / 0.48	0.58 / 0.56	0.53 / 0.50	0.53 / 0.51	0.37 / 0.48	0.45 / 0.53	0.72 / 0.61	0.64 / 0.59
TruFor	0.53 / 0.76	0.74 / 0.81	0.88 / 0.96	0.65 / 0.72	0.70 / 0.81	0.94 / 0.99	0.89 / 0.95	0.84 / 0.82	0.81 / 0.73
MVSS-Net	0.48 / 0.48	0.47 / 0.46	0.56 / 0.65	0.53 / 0.55	0.51 / 0.54	0.46 / 0.47	**0.88** / **0.97**	0.26 / 0.33	0.29 / 0.45
CAT-Net	0.57 / 0.52	0.46 / 0.37	0.85 / 0.91	0.85 / 0.91	0.68 / 0.68	**0.99** / **1.00**	0.72 / 0.83	0.76 / 0.49	0.73 / 0.41
HiFi-IFDL	0.50 / 0.50	0.47 / 0.35	0.51 / 0.52	0.51 / 0.53	0.49 / 0.47	0.50 / 0.51	0.50 / 0.55	0.21 / 0.33	0.21 / 0.54
NPR	0.50 / 0.51	0.90 / 0.98	0.50 / 0.50	0.52 / 0.52	0.60 / 0.63	0.50 / 0.49	0.51 / 0.52	0.92 / **0.97**	**0.94** / **0.98**
UFD	0.51 / 0.54	0.71 / 0.73	0.56 / 0.65	0.51 / 0.53	0.57 / 0.61	0.47 / 0.48	0.49 / 0.50	0.79 / 0.74	0.79 / 0.57
DE-FAKE	0.51 / 0.52	0.88 / 0.95	0.51 / 0.53	0.51 / 0.57	0.60 / 0.64	0.52 / 0.56	0.49 / 0.55	0.86 / 0.98	0.83 / 0.92
Ours	**0.86** / **0.88**	**0.93** / **0.94**	**0.95** / **0.99**	**0.93** / **0.97**	**0.92** / **0.94**	0.92 / 0.96	0.85 / 0.93	0.92 / 0.96	**0.94** / 0.96

7.1 Fake & Manipulated Image Detection

Setup. We benchmark our FKG system and competing approaches, including both forensic systems and modern vision-language models (VLMs), on our FKG-50K test set and four public datasets: DSO-1 [16], CASIA-v2 [22], GenImage [90], and Synthbuster [6]. We report overall and category-wise performance across forgery types: splicing, traditional editing, AI-editing, and fully AI-generated. VLM prompt templates and protocols are detailed in the appendix.

Metrics. We report detection accuracy and area under the ROC curve (AUC).

Competing Methods. We compare against forensic systems purpose-built for forgery detection (TruFor [34], MVSS-Net [18], CAT-Net [43], HiFi-IFDL [36] for traditional manipulations; NPR [74], UFD [57], DE-FAKE [69] for AI-generated detection) and powerful SOTA vision-language models (VLMs) including GPT-5 [60], Sonnet 4 [3], Gemini 2.0 Flash [31], Qwen3 VL [5], Gemma 3 [29], Llama 4 Scout [52], and Nemotron Nano [21].

Results. Tab. 1 reports detection accuracy and AUC across forgery types and datasets. On FKG-50K, we achieve the best overall accuracy (0.92) and AUC (0.94) with consistently strong per-category performance ($\geq$ 0.94 AUC except AI-edits at 0.88). On OOD datasets, we generalize well and remain competitive with the best specialized systems (DSO-1: 0.96, Synthbuster: 0.96).

Forensic systems only perform well on the specific forgery type they were trained for. Splicing detectors such as CAT-Net excel on their target domain (1.00 AUC on DSO-1) but fail on AI-generated content (0.37 AUC on Full AI), while AI-generation detectors like NPR show the opposite pattern (0.98 AUC on Synthbuster but 0.50 AUC on splicing). No single forensic system covers all forgery types. Conversely, VLMs can detect fully AI-generated images (GPT-5: 0.93 accuracy on GenImage, Qwen3 VL: 0.92 on Synthbuster) but fail on local manipulations (GPT-5: 0.59 on AI-edits, 0.55 on traditional edits). This suggests VLMs identify wholesale generation via image-wide artifacts such as

Table 2: Forgery type identification accuracy on FKG-50K and OOD datasets. VLMs show both real-world and oracle-assisted (in parentheses) scenarios. [I] = Instruct, [T] = Thinking.

	Method	FKG-50K					Out of Distribution Datasets			
							Splice & Edit		Fully AI-Gen	
		AI-Edit	Full AI	Splicing	Trad-Edit	Overall	DSO-1	CASIAv2	GenImage	Synthbuster
VLM	GPT-5	0.03 (0.02)	0.29 (0.39)	0.49 (0.75)	0.12 (0.56)	0.23 (0.43)	0.00 (0.79)	0.39 (0.81)	0.62 (0.54)	0.56 (0.57)
	Sonnet 4	0.04 (0.10)	0.44 (0.80)	0.04 (0.32)	0.02 (0.71)	0.14 (0.48)	0.00 (0.26)	0.03 (0.32)	0.26 (0.72)	0.55 (0.86)
	Gemini 2.0	0.03 (0.01)	0.23 (0.40)	0.40 (0.54)	0.04 (0.91)	0.18 (0.47)	0.04 (0.09)	0.22 (0.34)	0.52 (0.66)	0.54 (0.61)
	Qwen3 VL [I]	0.03 (0.37)	0.19 (0.13)	0.10 (0.16)	0.02 (0.77)	0.09 (0.36)	0.00 (0.00)	0.01 (0.00)	0.24 (0.09)	0.51 (0.39)
	Qwen3 VL [T]	0.45 (0.49)	0.45 (0.40)	0.29 (0.34)	0.05 (0.28)	0.31 (0.38)	0.00 (0.70)	0.06 (0.23)	0.38 (0.37)	0.65 (0.63)
	Gemma 3	0.13 (0.95)	0.28 (0.35)	0.25 (0.10)	0.21 (0.05)	0.22 (0.36)	0.07 (0.14)	0.25 (0.38)	0.38 (0.51)	0.50 (0.72)
	Llama 4 S.	0.04 (0.06)	0.15 (0.91)	0.02 (0.01)	0.16 (0.01)	0.09 (0.25)	0.00 (0.00)	0.05 (0.20)	0.16 (0.09)	0.21 (0.08)
	Nemotron N.	0.17 (0.25)	0.06 (0.33)	0.12 (0.40)	0.11 (0.43)	0.12 (0.35)	0.14 (0.56)	0.30 (0.31)	0.06 (0.20)	0.05 (0.21)
FKG	**Ours**	**0.81**	**0.84**	**0.92**	**0.89**	**0.87**	0.67	0.71	**0.91**	**0.94**

texture regularity or lighting inconsistencies [75, 80], but cannot detect localized manipulations that preserve overall visual coherence.

In contrast, our FKG generalizes across all forgery types because our backbone's novel self-supervised learning approach learns intrinsic forensic fingerprints of any imaging pipeline, AI generators included. This empowers the FRPN to detect any manipulation by identifying regions with differing forensic fingerprints, and task experts to detect fully AI-generated images via fingerprints distinct from any camera source. AI-edits are the hardest because AI editors condition on real pixels, generating content that partially inherits the original's forensic properties. Despite this, our FKG achieves 0.88 AUC, substantially outperforming the best forensic system (TruFor, 0.76) and best VLM (GPT-5, 0.60).

7.2 Forgery Type & Location Identification

Setup. This experiment tests forgery type identification and localization under two settings. Scenario 1 simulates real-world use, requiring simultaneous detection, classification, and localization, with missed detections penalized. Scenario 2 gives competitors a perfect oracle detector, while *our FKG operates identically in both*. All methods are evaluated on FKG-50K and public OOD datasets.

Metrics. We report forgery-type classification accuracy and localization F1 (predicted vs. ground-truth mask), following prior practices [18, 34]. VLM descriptions are converted to masks using Grounding Dino [48].

Competing Methods. Same as Sec. 7.1, excluding synthetic image detectors from Tabs. 2 and 3 (not designed to classify or localize forgeries), and forensic systems from Tab. 2 (not designed to classify forgery types).

Forgery Type Classification Results. Tab. 2 reports forgery type classification accuracy under real-world and oracle-assisted scenarios, where our FKG uses no oracle. On FKG-50K, we achieve the best overall type acc. (0.87) with strong per-category performance (0.81–0.92 across all forgery types). On OOD datasets, we maintain strong generalization (Synthbuster: 0.94, CASIAv2: 0.71).

VLMs fail at this task even under favorable conditions. On FKG-50K without oracle, the best VLM achieves only 0.31 type acc. (Qwen3 VL). With oracle, the best improves to just 0.48 (Sonnet 4), far below our 0.87. In rare instances, VLMs achieved strong performance with an oracle (GPT-5: 0.75 on splicing), though these gains are inconsistent and oracle access can even hurt (GPT-5: 0.62 $\rightarrow$

Table 3: Forgery localization F1 on FKG-50K and OOD datasets. VLMs show both real-world and oracle-assisted (in parentheses) scenarios. [I] = Instruct, [T] = Thinking.

	Method	FKG-50K				OOD	
						Splice & Edit	
		AI-Edit	Splicing	Trad-Edit	Overall	DSO-1	CASIAv2
VLM	GPT-5	0.17 (0.66)	0.59 (0.84)	0.10 (0.56)	0.29 (0.69)	0.00 (0.63)	0.43 (0.77)
	Sonnet 4	0.09 (0.46)	0.08 (0.47)	0.02 (0.53)	0.06 (0.49)	0.00 (0.43)	0.07 (0.47)
	Gemini 2.0	0.05 (0.50)	0.40 (0.72)	0.07 (0.45)	0.17 (0.56)	0.02 (0.34)	0.28 (0.65)
	Qwen3 VL [I]	0.05 (0.65)	0.25 (0.71)	0.05 (0.56)	0.12 (0.64)	0.00 (0.39)	0.25 (0.78)
	Qwen3 VL [T]	0.30 (0.50)	0.58 (0.69)	0.30 (0.50)	0.39 (0.56)	0.00 (0.34)	0.40 (0.61)
	Gemma 3	0.20 (0.36)	0.28 (0.48)	0.22 (0.48)	0.23 (0.44)	0.12 (0.22)	0.22 (0.40)
	Llama 4 S.	0.03 (0.02)	0.01 (0.03)	0.01 (0.03)	0.02 (0.03)	0.00 (0.10)	0.00 (0.05)
	Nemotron N.	0.17 (0.39)	0.18 (0.40)	0.13 (0.34)	0.16 (0.38)	0.15 (0.60)	0.22 (0.49)
Forensic Systems	TruFor	0.27	0.81	0.29	0.46	0.93	**0.83**
	MVSS-Net	0.32	0.59	0.22	0.38	0.25	0.63
	CAT-Net	0.38	0.64	0.40	0.47	0.59	0.66
	HiFi-IFDL	0.10	0.18	0.12	0.13	0.27	0.26
FKG	Ours	**0.93**	**0.98**	**0.91**	**0.94**	**0.95**	0.73

0.54 on GenImage). On OOD datasets, oracle-assisted VLMs show competitive numbers (GPT-5: 0.79 on DSO-1, 0.81 on CASIAv2). Without oracle, they score near zero while our FKG achieves its OOD performance unaided.

Localization Results. Tab. 3 reports forgery localization F1. On FKG-50K, we achieve the best overall F1 (0.94) with consistent per-category performance (0.91–0.98). On OOD data, we achieve 0.95 F1 on DSO-1.

Similar to Sec. 7.1, forensic systems only perform well on the forgery types they were trained for. For instance, TruFor excels on splicing (CASIAv2: 0.83 vs. our 0.73, DSO-1: 0.93 vs. our 0.95) but fails on AI-edits (0.27) and traditional edits (0.29). Similarly, VLMs without oracle cannot localize forgeries (best: 0.39). With oracle, some show reasonable results on specific categories (GPT-5: 0.84 on splicing, Qwen3 VL: 0.78 on CASIAv2). However, these gains are inconsistent and oracle access is unavailable in practice.

Two factors drive our strong performance across both tasks. First, accurate localization is enabled by the FRPN's novel Hybrid Graph Attention Transformer, which combines global self-attention to capture long-range forensic similarities with local graph attention to suppress false alarms from distant patches, yielding precise region boundaries. Second, accurate type classification is enabled by the task expert networks and the FKG ontology, which enables reasoning over cross-region relations: differing camera sources reveal splicing, absent sources reveal full synthesis, and mixed real/synthetic traces reveal AI-editing.

7.3 Forensic Decision Justification

Setup. This experiment measures each system's ability to generate forensic justifications that accurately and completely reference supporting evidence. Our FKG operates under real-world conditions, performing all tasks end-to-end, while competing VLMs are tested with oracle access to perfect ground-truth labels, isolating their reasoning abilities. All methods are evaluated on FKG-50K, which provides reference FKGs with full provenance histories.

Metrics. We assess justification quality with two metrics: correctness (whether forensic facts are described accurately) and completeness (whether all facts from the ground-truth FKG appear in the justification), averaged per image (Eq. (4)).

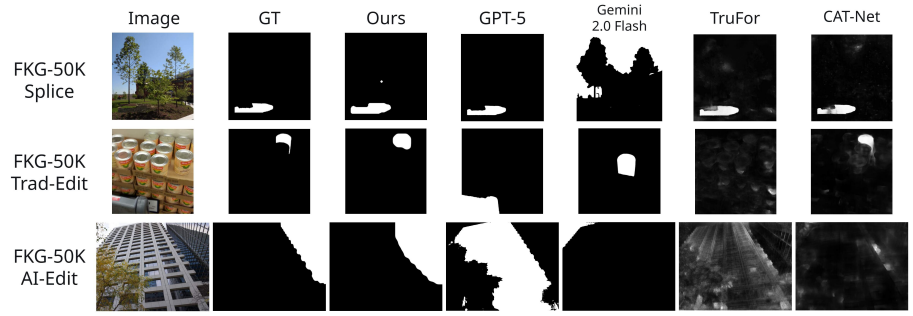

Fig. 6: Qualitative comparison of image forgery localization results.

Competing Methods. Because classical forensic models do not produce textual explanations, we compare only against the list of advanced VLMs capable of multimodal reasoning in Secs. 7.1 and 7.2.

Results. Tab. 4 reports forensic justification correctness and completeness on FKG-50K. Our FKG achieves the best overall correctness (0.75) and completeness (0.85), outperforming all VLMs despite operating without oracle access. Even with oracle access, VLMs reach only 0.16–0.24 correctness and 0.06–0.10 completeness. Per-category, we are strongest on full synthesis (0.86/0.96) and splicing (0.81/0.92), and lowest on traditional edits (0.65/0.75).

Table 4: Forensic justification correctness (COR) and completeness (COM) on FKG-50K.

	Method	AI-Edit		Full AI		Splice		Trad		Overall	
		COR	COM	COR	COM	COR	COM	COR	COM	COR	COM
VLM	GPT-5	0.15	0.05	0.28	0.16	0.33	0.12	0.08	0.03	0.21	0.09
	Sonnet 4	0.17	0.05	0.37	0.27	0.16	0.05	0.05	0.03	0.19	0.10
	Gemini 2.0	0.17	0.02	0.30	0.14	0.38	0.06	0.10	0.03	0.24	0.06
	Qwen3 VL [I]	0.25	0.08	0.05	0.05	0.26	0.08	0.15	0.06	0.18	0.07
	Qwen3 VL [T]	0.28	0.09	0.20	0.14	0.30	0.09	0.11	0.04	0.22	0.09
	Gemma 3	0.36	0.10	0.28	0.14	0.24	0.07	0.06	0.02	0.23	0.08
	Llama 4 S.	0.05	0.01	0.62	0.28	0.03	0.01	0.00	0.00	0.17	0.08
	Nemotron N.	0.21	0.08	0.18	0.13	0.18	0.08	0.07	0.03	0.16	0.08
FKG	**Ours**	**0.69**	**0.77**	**0.86**	**0.96**	**0.81**	**0.92**	**0.65**	**0.75**	**0.75**	**0.85**

These results extend the findings from Secs. 7.1 and 7.2: even when given the forgery type and exact location, VLMs cannot explain the forensic evidence behind a manipulation. The completeness gap is particularly revealing: VLMs cover at most 10% of forensic facts (Sonnet 4: 0.10), compared to our 85%. Their justifications rely on visual speculation ("the lighting looks inconsistent") rather than verifiable forensic traces such as camera model discrepancies, compression inconsistencies, or modification lineage.

Our FKG produces faithful justifications because the FKG ontology encodes each region's source identity, compression history, and modification lineage as structured triplets, giving the VLM (Sec. 5) concrete facts to report rather than requiring it to infer forensic conclusions from pixels. Iterative Context Refinement then optimizes the VLM's in-context examples to maximize both correctness and completeness without fine-tuning.

8 Discussion

Effects of Iterative Context Refinement. We compare the forensic justification quality of different prompting strategies (Tab. 5) on identical FKG inputs

from FKG-50K. The results show that our ICR protocol achieves the highest correctness (0.75) and completeness (0.85). Without any context, the VLM covers only 32% of forensic facts. Random in-context examples improve completeness to 0.51 by demonstrating the expected output format, and adding CoT [82] further improves correctness to 0.70. However, the largest gain comes from ICR, which nearly doubles completeness from 0.55 to 0.85 by iteratively selecting demonstrations that target the VLM's specific failure patterns. Crucially, ICR achieves this without any VLM fine-tuning, operating purely at the prompt level.

FKG Accuracy and Failure Modes. Beyond the correctness and completeness reported in Sec. 7.3, we evaluate FKG generation accuracy using *perfect match*, which requires every fact in the predicted FKG to exactly match the ground truth. Our system achieves 0.65 perfect match overall, with the highest accuracy on full synthesis (0.81) and splicing (0.68), and the lowest on AI-edits (0.58) and traditional edits (0.52).

Table 5: Effect of optimization strategies on forensic justification quality.

Opt. Strategy	CoT	Correctness	Completeness
None	X	0.64	0.32
Random	X	0.63	0.51
Random	✓	0.70	0.55
ICR protocol	✓	**0.75**	**0.85**

The gap between correctness (0.75) and perfect match (0.65) reveals that most errors are in fine-grained parameters (*e.g.*, generator model, compression quality) rather than structural relationships, which the system consistently recovers. Full synthesis achieves the highest perfect match (0.81) because its FKGs are structurally simple: one region, one AI source, with fewer parameters to predict. Splicing follows (0.68) due to sharp boundaries between clearly distinct sources. Traditional edits are hardest (0.52) because operations like blur and resampling leave subtle traces that make exact parameter recovery difficult. AI-edits present a different challenge (0.58): some AI generators produce similar forensic fingerprints, making exact source attribution difficult for the current task expert. Fig. 6 shows representative success and failure cases.

Despite these parameter-level errors, authentication performance remains strong (Sec. 7) because correct graph topology suffices for detection, localization, and justification. This structural robustness also explains out-of-distribution generalization: the self-supervised backbone learns forensic fingerprints from imaging pipeline properties rather than image content, and the ontology structures evidence as universal forensic properties rather than forgery-specific features.

Human Trust Evaluation. A core premise of our framework is that grounding authentication in structured forensic evidence makes its decisions more trustworthy to users. To test this directly, 30 participants rated their trust (1 to 5) in our system's output across 5 images, each shown both as a bare verdict and with its accompanying forensic graph explanation. Presenting the graph raised average trust from 3.39 to 4.53, confirming that auditable, evidence-grounded explanations substantially increase user trust over a verdict alone.

Limitations. While our framework provides a strong foundation for trustworthy image authentication, fine-grained parameter prediction and error propagation from imperfect region proposals remain the primary bottlenecks for perfect FKG reconstruction. Importantly, both are addressable through the framework's

modular design: more accurate task experts or region proposal methods can be integrated without modifying the ontology, the reasoning pipeline, or ICR. Extended analysis, including computational costs, is provided in the appendix.

9 Ablation

We conduct comprehensive ablation studies of each major system component on FKG-50K, shown in Tab. 6.

No self-supervision. Without self-supervised fingerprint learning, the backbone relies only on task labels and loses its ability to encode generalizable forensic microstructures, causing all metrics to drop sharply.

No pairwise similarity. Keeping only the contrastive loss removes the pairwise similarity constraint, weakening structural alignment, reducing type and localization accuracy. Contrastive learning alone fails to capture relational consistency among patches.

Table 6: Ablation study of system components on FKG-50K.

Stage	Variant	Detect AUC	Type ACC	Loc F1
–	Ours	**0.94**	**0.87**	**0.94**
Backbone	No self-supervision	0.72 (-0.22)	0.66 (-0.21)	0.44 (-0.50)
	No pairwise similarity	0.90 (-0.04)	0.63 (-0.24)	0.86 (-0.08)
FRPN	Only local graph-attention	0.85 (-0.09)	0.57 (-0.30)	0.93 (-0.01)
	Only full self-attention	0.92 (-0.02)	0.55 (-0.32)	0.93 (-0.01)
	Only contrastive loss	0.89 (-0.05)	0.52 (-0.35)	0.72 (-0.22)
	Only Hungarian loss	0.81 (-0.13)	0.45 (-0.42)	0.59 (-0.35)
Attr. Infer	No transformer reasoner	0.94 (±0.00)	0.69 (-0.18)	0.94 (±0.00)

Only local graph-attention. Disabling global self-attention limits reasoning to local neighborhoods, causing large drops in type accuracy. Long-range dependencies are essential to model global forgery context.

Only full self-attention. Removing local graph attention breaks spatial coherence, sharply degrading type accuracy. Local connectivity is necessary to preserve region homogeneity and prevent attention dilution.

Only contrastive loss. Training with only contrastive loss omits Hungarian matching, yielding unstable and inconsistent region assignments. Contrastive learning alone cannot form discrete, semantically aligned regions.

Only Hungarian loss. Eliminating the contrastive term leaves only discrete alignment, reducing detection and classification accuracy. Continuous similarity constraints are crucial for coherent region grouping.

No transformer reasoner. Replacing the transformer reasoner with linear classifiers drops type accuracy by 0.18. The reasoning module is essential for integrating cross-region evidence to infer manipulation semantics.

10 Conclusion

We introduced Forensic Knowledge Graphs, a unified framework for trustworthy image authentication that integrates forensic evidence extraction, structured reasoning, and interpretable explanation. By modeling images as structured graphs linking regions, sources, and manipulations, our system delivers accurate, localized, and explainable authenticity judgments across diverse forgery types.

References

1. Amerini, I., Barni, M., Battiato, S., Bestagini, P., Boato, G., Bonaventura, T.S., Bruni, V., Caldelli, R., De Natale, F., De Nicola, R., et al.: Deepfake media forensics: State of the art and challenges ahead. arXiv preprint arXiv:2408.00388 (2024)
2. Anlen, L., Vazquez Llorente, R.: Spotting the deepfakes: how AI detection tools work and where they fail. `https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail` (2024), reuters Institute / WITNESS, accessed: 2026-06-25
3. Anthropic: System card: Claude Opus 4 & Claude Sonnet 4. `https://www.anthropic.com/claude-4-system-card` (2025)
4. Bagaria, A.: INSIGHT: An interpretable neural vision-language framework for reasoning of generative artifacts. arXiv preprint arXiv:2511.22351 (2025)
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
6. Bammey, Q.: Synthbuster: Towards detection of diffusion model generated images. IEEE Open Journal of Signal Processing **5**, 1–9 (2024)
7. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020)
8. Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: FLUX.1 kontext: Flow matching for in-context image generation and editing in latent space. arXiv preprint arXiv:2506.15742 (2025)
9. Bayar, B., Stamm, M.C.: A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security. pp. 5–10 (2016)
10. Bayar, B., Stamm, M.C.: Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. IEEE Transactions on Information Forensics and Security **13**(11), 2691–2706 (2018)
11. Black Forest Labs: FLUX.1: A suite of text-to-image generation models. `https://blackforestlabs.ai` (2024)
12. Brooks, T., Holynski, A., Efros, A.A.: InstructPix2Pix: Learning to follow image editing instructions. In: CVPR. pp. 18392–18402 (2023)
13. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS. vol. 33, pp. 1877–1901 (2020)
14. Cao, H., Mei, Q., Li, Z., Li, Y., Meng, Z., Zhang, Y., Li, C., Zhang, Z., Ding, X., Wang, Y., Lyu, J., Wu, F.: REVEAL: Reasoning-enhanced forensic evidence analysis for explainable AI-generated image detection. arXiv preprint arXiv:2511.23158 (2025)
15. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020)
16. de Carvalho, T.J., Riess, C., Angelopoulou, E., Pedrini, H., de Rezende Rocha, A.: Exposing digital image forgeries by illumination color classification. IEEE Transactions on Information Forensics and Security **8**(7), 1182–1194 (2013)
17. Caulfield, M.: How to use ChatGPT to detect AI (and otherwise digitally altered) photos. `https://mikecaulfield.substack.com/p/how-to-use-chatgpt-to-detect-ai-and` (2025), accessed: 2026-06-25

18. Chen, X., Dong, C., Ji, J., Cao, J., Li, X.: Image manipulation detection by multi-view multi-scale supervision. In: ICCV. pp. 14185–14193 (2021)
19. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: ICASSP. pp. 1–5 (2023)
20. Cozzolino, D., Verdoliva, L.: Noiseprint: A CNN-based camera model fingerprint. IEEE Transactions on Information Forensics and Security **15**, 144–159 (2020)
21. Deshmukh, A.S., Chumachenko, K., Rintamaki, T., Le, M., Poon, T., et al.: NVIDIA Nemotron Nano V2 VL. arXiv preprint arXiv:2511.03929 (2025)
22. Dong, J., Wang, W., Tan, T.: CASIA image tampering detection evaluation database. In: IEEE China Summit and International Conference on Signal and Information Processing. pp. 422–426 (2013)
23. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
24. Edwards, L., Wojciak, T., Anlen, L.: AI is undermining OSINT's core assumptions: here's how journalists should adapt. `https://reutersinstitute.politics.ox.ac.uk/news/ai-undermining-osints-core-assumptions-heres-how-journalists-should-adapt` (2025), reuters Institute / WITNESS, accessed: 2026-06-25
25. Europol Innovation Lab: Facing reality? Law enforcement and the challenge of deepfakes. Tech. rep., Europol (2022)
26. Full Fact: Grok and Google Lens AI overviews claim fake imagery shows Huntingdon train attack. `https://fullfact.org/crime/grok-google-lens-ai-imagery-train-attack/` (2025), accessed: 2026
27. Gemini Team: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
28. Gemini Team, Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
29. Gemma Team, Kamath, A., Ferret, J., Pathak, S., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
30. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
31. Google DeepMind: Gemini 2.0 Flash model card. `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-0-Flash-Model-Card.pdf` (2025), accessed: 2026-06-25
32. Groh, M., Epstein, Z., Firestone, C., Picard, R.: Deepfake detection by human crowds, machines, and machine-informed crowds. Proceedings of the National Academy of Sciences **119**(1), e2110013119 (2022)
33. Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition **5**(2), 199–220 (1993)
34. Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., Verdoliva, L.: TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In: CVPR. pp. 20606–20615 (2023)
35. Gunning, D., Aha, D.W.: DARPA's explainable artificial intelligence (XAI) program. AI Magazine **40**(2), 44–58 (2019)
36. Guo, X., Liu, X., Ren, Z., Grosz, S., Masi, I., Liu, X.: Hierarchical fine-grained image forgery detection and localization. In: CVPR. pp. 3155–3165 (2023)

37. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV. pp. 2961–2969 (2017)
38. Huang, Z., Hu, J., Li, X., He, Y., Zhao, X., Peng, B., Wu, B., Huang, X., Cheng, G.: SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In: CVPR (2025)
39. Jia, S., Lyu, R., Zhao, K., Chen, Y., Yan, Z., Ju, Y., Hu, C., Li, X., Wu, B., Lyu, S.: Can ChatGPT detect DeepFakes? A study of using multimodal large language models for media forensics. In: CVPRW. pp. 4324–4333 (2024)
40. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: CVPR. pp. 8107–8116 (2020)
41. Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T.T., Mober, H., et al.: DSPy: Compiling declarative language model calls into state-of-the-art pipelines. In: ICLR (2024)
42. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
43. Kwon, M.J., Yu, I.J., Nam, S.H., Lee, H.K.: CAT-Net: Compression artifact tracing network for detection and localization of image splicing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 375–384 (2021)
44. Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., Zhou, B.: Trustworthy AI: From principles to practices. ACM Computing Surveys **55**(9), 1–46 (2023)
45. Li, X., Lv, K., Yan, H., Lin, T., Zhu, W., Ni, Y., Xie, G., Wang, X., Qiu, X.: Unified demonstration retriever for in-context learning. In: ACL. pp. 4644–4668 (2023)
46. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
47. Liu, J., Zhang, F., Zhu, J., Sun, E., Zhang, Q., Zha, Z.J.: ForgeryGPT: Multimodal large language model for explainable image forgery detection and localization. arXiv preprint arXiv:2410.10238 (2024)
48. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: ECCV (2024)
49. Lukas, J., Fridrich, J., Goljan, M.: Digital camera identification from sensor pattern noise. IEEE Transactions on Information Forensics and Security **1**(2), 205–214 (2006)
50. Mareen, H., Karageorgiou, D., Van Wallendael, G., Lambert, P., Papadopoulos, S.: TGIF: Text-guided inpainting forgery dataset. In: IEEE International Workshop on Information Forensics and Security (WIFS). pp. 1–6 (2024)
51. Mayer, O., Stamm, M.C.: Forensic similarity for digital images. IEEE Transactions on Information Forensics and Security **15**, 1331–1346 (2020)
52. Meta AI: The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/` (2025)
53. Midjourney: Midjourney. `https://www.midjourney.com`, accessed: 2025
54. Nguyen, T.D., Azizpour, A., Stamm, M.C.: Forensic self-descriptions are all you need for zero-shot detection, open-set source attribution, and clustering of AI-generated images. In: CVPR. pp. 3040–3050 (2025)
55. Nightingale, S.J., Farid, H.: AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proceedings of the National Academy of Sciences **119**(8), e2120481119 (2022)

56. Nightingale, S.J., Wade, K.A., Watson, D.G.: Can people identify original and manipulated photos of real-world scenes? Cognitive Research: Principles and Implications **2**(1), 30 (2017)
57. Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: CVPR. pp. 24480–24489 (2023)
58. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
59. OpenAI: GPT-Image-1: Native image generation in ChatGPT. `https://openai.com/index/introducing-4o-image-generation/` (2025)
60. OpenAI: OpenAI GPT-5 system card. arXiv preprint arXiv:2601.03267 (2025)
61. Pevny, T., Fridrich, J.: Detection of double-compression in JPEG images for applications in steganography. IEEE Transactions on Information Forensics and Security **3**(2), 247–258 (2008)
62. Pexels: Pexels: Free stock photos, royalty free images and videos. `https://www.pexels.com`, accessed: 2025
63. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
64. Pryzant, R., Iter, D., Li, J., Lee, Y.T., Zhu, C., Zeng, M.: Automatic prompt optimization with "gradient descent" and beam search. In: EMNLP. pp. 7957–7968 (2023)
65. Qwen Team: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
66. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
67. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
68. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to detect manipulated facial images. In: ICCV. pp. 1–11 (2019)
69. Sha, Z., Li, Z., Yu, N., Zhang, Y.: DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. pp. 3418–3432 (2023)
70. Stamm, M.C., Liu, K.J.R.: Forensic detection of image manipulation using statistical intrinsic fingerprints. IEEE Transactions on Information Forensics and Security **5**(3), 492–506 (2010)
71. Stamm, M.C., Liu, K.J.R.: Anti-forensics of digital image compression. IEEE Transactions on Information Forensics and Security **6**(3), 1050–1065 (2011)
72. Stamm, M.C., Wu, M., Liu, K.J.R.: Information forensics: An overview of the first decade. IEEE Access **1**, 167–200 (2013)
73. Tan, C., Wang, J., Ming, X., Tao, R., Wei, Y., Zhao, Y., Lu, Y.: ForenX: Towards explainable AI-generated image detection with multimodal large language models. arXiv preprint arXiv:2508.01402 (2025)
74. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. In: CVPR. pp. 28130–28139 (2024)
75. Tariq, S., Nguyen, D., Chamikara, M., Wu, T., Abuadbba, A., Moore, K.: LLMs are not yet ready for deepfake image detection. arXiv preprint arXiv:2506.10474 (2025)
76. Unsplash: Unsplash: Beautiful free images and pictures. `https://unsplash.com`, accessed: 2025

77. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
78. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph attention networks. In: ICLR (2018)
79. Verdoliva, L.: Media forensics and deepfakes: an overview. IEEE Journal of Selected Topics in Signal Processing **14**(5), 910–932 (2020)
80. Wang, J., Lv, C., Li, X., Dong, S., Li, H., Yao, K., Li, C., Shao, W., Luo, P.: Forensics-Bench: A comprehensive forgery detection benchmark suite for large vision language models. In: CVPR (2025)
81. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-generated images are surprisingly easy to spot... for now. In: CVPR (2020)
82. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS (2022)
83. Xu, Z., Zhang, X., Li, R., Tang, Z., Huang, Q., Zhang, J.: FakeShield: Explainable image forgery detection and localization via multi-modal large language models. In: ICLR (2025)
84. Yan, S., Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Xie, W.: A sanity check for AI-generated image detection. In: ICLR (2025)
85. Yang, J., Zhang, H., Li, F., Zou, X., Li, C., Gao, J.: Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. arXiv preprint arXiv:2310.11441 (2023)
86. Ye, J., Wu, Z., Feng, J., Yu, T., Kong, L.: Compositional exemplars for in-context learning. In: ICML. pp. 39818–39833 (2023)
87. Zhao, X., Stamm, M.C.: Computationally efficient demosaicing filter estimation for forensic camera model identification. In: ICIP. pp. 151–155 (2016)
88. Zheng, C., Lin, C., Zhao, Z., Wang, H., Guo, X., Liu, S., Shen, C.: Breaking semantic artifacts for generalized AI-generated image detection. In: NeurIPS (2024)
89. Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., et al.: Judging LLM-as-a-judge with MT-Bench and chatbot arena. In: NeurIPS (2023)
90. Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y.: GenImage: A million-scale benchmark for detecting AI-generated image. In: NeurIPS (2023)