EFG: ENCODER-FREE FORENSIC GUARD FOR SYNTHETIC IMAGE DETECTION AND EXPLANATION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image detection has recently benefited from vision-language models (VLMs), which provide natural-language explanations to justify their judgments. However, existing VLM-based detectors heavily rely on visual encoders, which compress images into high-level semantic features, inevitably discarding low-level synthetic artifacts critical for forensic discrimination. Consequently, this compression creates a fundamental bottleneck, with detection-relevant signals largely lost before the model even begins reasoning. In this work, we propose EFG (Encoder-Free Forensic Guard), a specialized VLM that captures comprehensive forensic signals without visual encoders. Specifically, EFG converts the input image into token sequences via a lightweight patch embedding layer and directly feeds these raw pixel-level tokens into the large language model. This design preserves complete visual signals without encoder-induced information loss, while adaptively fusing low-level forensic traces with high-level semantics across layers. Beyond detection, EFG can produce natural-language explanations to justify its forensic judgments. Extensive experiments show that EFG achieves state-of-the-art performance across three AI-generated image detection benchmarks. Our work highlights the potential of encoder-free, pixel-space learning for VLM-based forensic systems. Code will be open-sourced upon acceptance.

1 INTRODUCTION

Recent advances in generative models, including GANs (Goodfellow et al., 2014), diffusion models (Ho et al., 2020), and large-scale multimodal generation systems (Zhou et al., 2025a), have enabled the creation of synthetic images that defy visual scrutiny, raising unprecedented challenges to digital trust and authenticity. Consequently, developing reliable detection techniques for AI-generated images has become an urgent imperative in the fields of digital forensics and content moderation. Although prior work (Wang et al., 2020; Zhong et al., 2024) has made notable progress in detecting synthetic imagery, these methods remain largely confined to binary decisions on authenticity, rendering a verdict without explanation. This lack of interpretability severely undermines both the adaptability and the transparency necessary for trustworthy practical applications.

The emergence of vision-language models (VLMs) has opened new avenues for explainable synthetic image detection. By leveraging their strong visual-linguistic reasoning capabilities, recent studies (Jia et al., 2024; Li et al., 2024) have explored the use of VLMs for synthetic image detection and explanation. A subsequent wave of research (Wen et al., 2025; Jiang et al., 2026) has focused on fine-tuning VLMs, adapting these general-purpose models into specialized forensic detectors, thereby achieving markedly improved performance. However, a critical limitation persists in existing VLM-based detectors: the visual encoders they rely on are inherently optimized for semantic content, which suppresses fine-grained artifacts and prevents the model from reasoning over these low-level cues. This semantic-centric design not only hinders detection but also undermines explainability, since the cues are discarded as noise by the visual encoder. To empirically verify this encoding bottleneck, we first fine-tune the visual encoder on Holmes-Set (Zhou et al., 2025b) prior to VLM fine-tuning and compare it against the frozen counterpart. As shown in Figure 1, the tuned-encoder VLM consistently outperforms the frozen one across both backbones, confirming the frozen encoder as a key obstacle to forensic detection. However, fine-tuning adds an extra stage without resolving forensic–semantic misalignment. This motivates us to re-examine the role of visual encoding toward more effective feature integration across forensic and semantic representations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we propose EFG (Encoder-Free Forensic Guard), a VLM-based framework for synthetic image detection and explanation. Unlike existing methods that rely on pretrained visual encoders, EFG adopts a lightweight patch embedding layer that directly converts raw images into token sequences fed into the large language model (LLM), bypassing encoder-induced information loss entirely. By removing the visual encoder entirely, our framework preserves the full spectrum of visual signals—from low-level pixel statistics to high-level semantic patterns—across all layers of the model. This encoder-free design enables EFG to reason over the very forensic cues that conventional encoders inevitably discard, while adaptively fusing these fine-grained traces with contextual semantics through the LLM's deep reasoning capacity. In contrast to two-stage approaches that fine-tune the visual encoder and the LLM backbone separately, EFG enables end-to-end training, where the entire model is jointly optimized toward the final forensic objective without task misalignment.

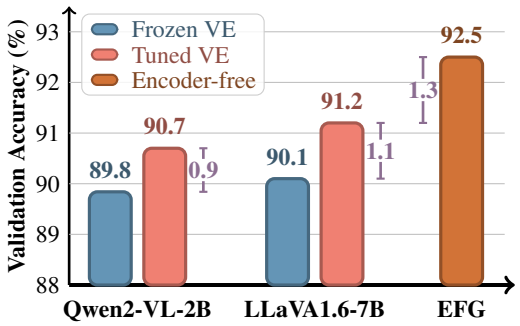

Figure 1: Detection accuracy on Holmes-Set for two VLMs using frozen vs. tuned visual encoders (VEs), alongside the proposed EFG approach.

Experimental results show that EFG achieves state-of-the-art performance on three benchmarks, surpassing previous specialized VLMs and conventional detectors. Specifically, EFG outperforms conventional detectors by 6.9% on FakeClue and by 5.9% on LOKI in detection accuracy. Our model also achieves superior performance in explanation quality, demonstrating the benefit of fusing low-level forensic traces with high-level contextual semantics. Additionally, EFG demonstrates robust zero-shot cross-dataset generalization on DMImage, further validating the effectiveness of our encoder-free design.

Our main contributions are threefold:

- We identify the visual encoder as a fundamental bottleneck in VLM-based forensic detection, showing that semantic prioritization directly undermines access to low-level forgery cues.
- We propose EFG, an encoder-free VLM tailored for synthetic image detection and explanation, which operates directly on raw pixels with no visual encoder.
- Extensive experiments demonstrate that EFG achieves state-of-the-art performance across multiple benchmarks, with strong generalizability to unseen generators and robust explanatory faithfulness.

2 Related Work

2.1 Discriminative Synthetic Image Detection

Conventional synthetic image detection methods (Wang et al., 2020; Zhu et al., 2023) predominantly framed the task as binary classification, employing convolutional or transformer backbones to capture spatial pixel anomalies (Zhong et al., 2024), frequency-domain irregularities (Tan et al., 2024b), reconstruction discrepancies (Wang et al., 2023), or other statistical artifacts (Bi et al., 2023a; Tan et al., 2024a). While these approaches achieve strong performance on standard benchmarks, their generalization across diverse real-world generative models remains limited. In response, DRCT (Chen et al., 2024a) and FSD (Wu et al., 2025) adopt contrastive learning objectives to enforce feature discriminability, while DDA (Chen et al., 2025a) and B-Free (Guillaro et al., 2025) focus on data rebalancing and debiasing, aiming to suppress biases originating from semantic content and lossy image compression. Recent methods have explored pre-trained models for better generalization. For instance, UnivFD (Ojha et al., 2023) and OOC-CLIP (Liu et al., 2024) leverage pre-trained CLIP-ViT (Radford et al., 2021) features to handle out-of-distribution data. C2P-CLIP (Tan et al., 2025) further injects category-related concepts to enhance feature discriminability. Despite their advances, these methods remain black-box and lack interpretability in their detection decisions.

2.2 Explainable Synthetic Detection

Recent advances in vision-language models (VLMs) enable interpretable synthetic image detection. Several studies (Li et al., 2024; Ye et al., 2025) have explored VLMs for synthetic image detection, where these models provide decisions and explanations, aligning more closely with human judgment. While off-the-shelf VLMs possess zero-shot detection capability, they still underperform conventional detectors on standard benchmarks. Thus, approaches such as FakeShield (Xu et al., 2025), SIDA (Huang et al., 2025), and FakeVLM (Wen et al., 2025) fine-tune VLMs on detection-specific data to obtain specialized detectors. Additionally, Ivy-xDetector (Jiang et al., 2026) has introduced chain-of-thought reasoning into the detection pipeline, while UniGenDet (Zhang et al., 2026) explores the co-evolutionary dynamics between image generation and detection within a unified framework. Other works have instead focused on adapting the visual encoder. AIGI-Holmes (Zhou et al., 2025b) fine-tunes the visual encoder and incorporates an established detector (Tan et al., 2024a) to enhance forgery features. Deep-VRM (Lin et al., 2026) extracts artifact signals from its low-level encoder and injects them into later layers, fusing with semantic tokens for multimodal reasoning. Despite these advances, effective fusion of forgery signals and semantic representations remains a major challenge.

2.3 Native Vision-Language Models

Native VLMs aim to achieve early-fusion integration for unified multimodal reasoning, contrasting with modular architectures that attach external visual encoders to large language models. Early efforts, including Fuyu (Bavishi et al., 2023), EVE (Diao et al., 2024), and SOLO (Chen et al., 2024b), directly embed image patches into the language space through linear projections, demonstrating the feasibility of native cross-modal integration. Other approaches (Chameleon Team, 2024; Zhou et al., 2025a) adopt discrete tokenization for visual inputs, treating both modalities as homogeneous streams, yet the modality gap remains a critical challenge. Recent studies (Beyer et al., 2024; Lei et al., 2025) show that with sufficient data and progressive optimization, native VLMs can rapidly approach modular counterparts. Subsequently, Mono-InternVL (Luo et al., 2025) and NEO (Diao et al., 2025) propose pre-training strategies that align multimodal content within a single transformer. With their strong capability for low-level visual perception, these models are inherently suited for fine-grained visual analysis. This makes native VLMs a promising candidate for detecting forensic cues, which we systematically explore in this work.

3 Method

In this section, we present EFG (Encoder-Free Forensic Guard), a native vision-language framework for AI-generated image detection and explanation. We first introduce the native visual embedding module, which directly maps raw image signals into visual tokens while preserving low-level forensic cues. Then we describe the backbone architecture for cross-modal processing. Finally, we present our parameter-efficient fine-tuning strategy to mitigate overfitting while preserving general reasoning capabilities. The overall pipeline is illustrated in Figure 2.

3.1 Native Visual Embedding

Most existing VLM-based forensic systems leverage a pre-trained visual backbone, such as CLIP-ViT (Radford et al., 2021), to extract visual features for semantic analysis. Given an input image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, the encoder processes it as:

$$\mathbf{x}_v = \text{CLIP-ViT}(\boldsymbol{I}) \in \mathbb{R}^{N\times d_v}, \tag{1}$$

where $N = HW/P^2$ (with P the patch size) denotes the number of patches, and d_v is the feature dimension. These visual features $\mathbf{x}_v$ are linearly projected to dimension d to align with the text token embeddings, then fed into the unified transformer backbone together with the text tokens. However, this encoder is inherently insensitive to low-level forensic cues due to its semantic-aligned optimization, leaving the subsequent transformer unable to access such essential signals. To address these limitations, we adopt a native VLM architecture that directly encodes the input image into visual tokens via a lightweight patch embedding layer. Specifically, we apply two convolutional

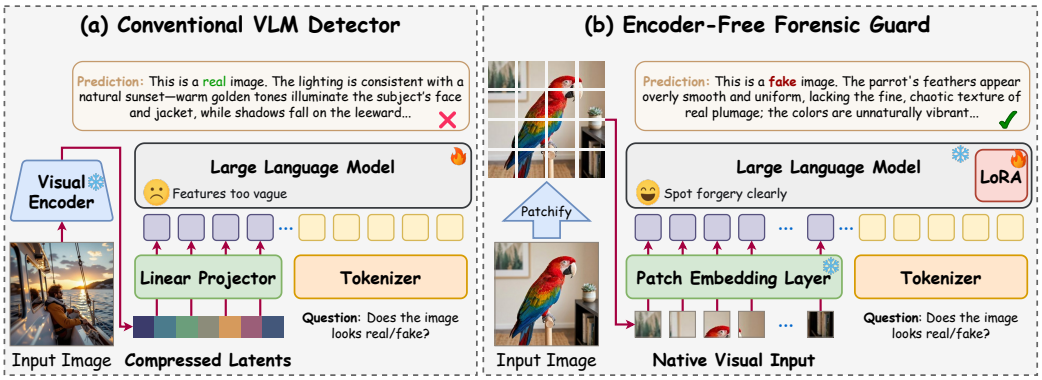

Figure 2: Comparison between (a) conventional VLM-based detectors (e.g., FakeVLM) with visual encoders and (b) our encoder-free EFG architecture. EFG directly processes raw image patches, preserving complete forensic signals for LLM reasoning.

layers with an activation function and positional encoding:

$$\mathbf{x}_v = \text{Conv}_2(\sigma(\text{Conv}_1(\boldsymbol{I})) + \mathbf{PE}) \in \mathbb{R}^{(h\times w)\times d}, \tag{2}$$

where σ denotes the activation function and **PE** is the 2D sinusoidal positional encoding (Dosovitskiy et al., 2021). The strides of Conv_1 and Conv_2 are set to 16 and 2, respectively. Conv_1 performs initial patch embedding, while Conv_2 performs token folding analogous to pixel unshuffle. Consequently, the output feature map has spatial dimensions $h = H/32$ and $w = W/32$, and is then flattened into a sequence of $h \times w$ visual tokens, each corresponding to a 32×32 image patch. Special tokens `<img>` and `</img>` are inserted at the boundaries of visual token sequences to demarcate visual modality boundaries. For textual input, we use the tokenizer of the backbone model to obtain word token embeddings $\mathbf{x}_t \in \mathbb{R}^{n\times d}$. The visual and textual token sequences are then concatenated and fed into the backbone, where cross-modal reasoning is performed natively without any modality-specific projection or adapter layers.

Our encoder-free design offers three key advantages over encoder-based approaches. First, by eliminating the pre-trained visual encoder, our model preserves raw image signals without compression or semantic re-projection, allowing low-level artifact cues to remain accessible throughout the network. Second, the lightweight patch embedding operates on the input image with only minimal resizing to align it to the 32×32 patch grid, avoiding the aggressive fixed-resolution preprocessing that would otherwise compromise low-level forensic traces. Third, our single-stage supervised fine-tuning enables the VLM to directly learn the correlation between raw artifact signals and semantic reasoning, eliminating the additional fine-tuning of a visual encoder required by prior work (Zhou et al., 2025b) and enabling end-to-end reasoning directly on native image signals.

3.2 Backbone Architecture

We adopt NEO (Diao et al., 2025) as our backbone, a native VLM that inherently establishes coherent pixel–word correspondences without relying on external vision encoders. NEO utilizes additional layers as a pre-buffer to bridge the native visual embeddings with the large language model (LLM) representation space, enabling seamless transfer of pre-trained linguistic capabilities to the multimodal setting. All tokens are processed by the same decoder layers without modality-specific branches or cross-attention mechanisms. Internally, each LLM layer processes the input sequence through multi-head attention (MHA) followed by a feed-forward network (FFN), and each module is preceded by RMSNorm (Zhang & Sennrich, 2019), equipped with residual connections:

$$\mathbf{x}_m^{l'} = \mathbf{x}_m^{l} + \text{MHA}(\text{RMSNorm}(\mathbf{x}_m^{l})), \tag{3}$$

$$\mathbf{x}_m^{l+1} = \mathbf{x}_m^{l'} + \text{FFN}(\text{RMSNorm}(\mathbf{x}_m^{l'})), \tag{4}$$

where $\mathbf{x}_m^l$ denotes the hidden representation of the m-th token at the l-th layer within a unified multimodal sequence, sharing the same computation.

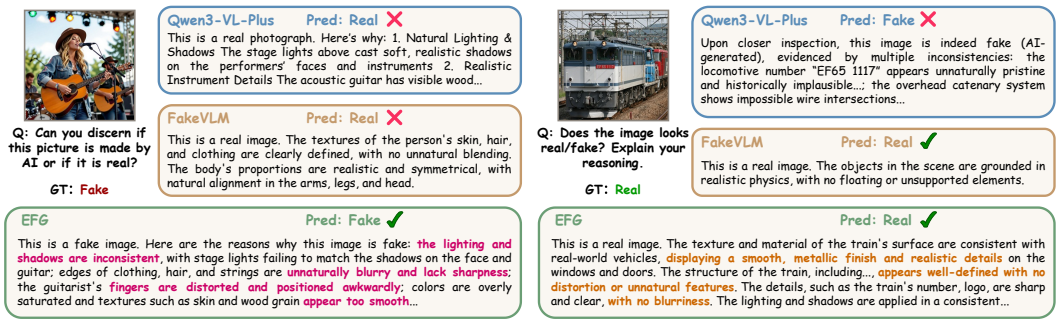

Figure 3: Qualitative comparison of explanation generation among Qwen3-VL-Plus, FakeVLM and EFG on representative samples, using official API or open-source weights. Key judgments of EFG are highlighted in bold.

During the supervised fine-tuning stage, the model is optimized by the standard autoregressive next-token prediction objective:

$$\mathcal{L}_{\text{SFT}} = -\sum_{i \in \mathcal{Y}} \log p(y_i \mid y_{<i}, \mathbf{c}), \tag{5}$$

where $\mathcal{Y}$ denotes the set of answer token positions, and $\mathbf{c}$ represents the preceding context, including both visual and textual tokens.

3.3 Parameter-Efficient Fine-Tuning

To efficiently adapt the pre-trained backbone to forensic detection tasks, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) instead of full-parameter fine-tuning. LoRA introduces trainable low-rank decomposition matrices to the attention layers while keeping the original model parameters frozen, significantly reducing the number of trainable parameters and mitigating the risk of overfitting when training on domain-specific forensic datasets. We apply LoRA to all linear projection matrices of the transformer layers in NEO, with rank r and scaling factor α. The adaptation is formulated as:

$$\mathbf{W} = \mathbf{W}_0 + \Delta\mathbf{W} = \mathbf{W}_0 + \mathbf{B}\mathbf{A}, \tag{6}$$

where $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$ represents the pre-trained weight matrix, $\mathbf{B} \in \mathbb{R}^{d\times r}$ and $\mathbf{A} \in \mathbb{R}^{r\times k}$ are the learnable low-rank matrices. During training, only $\mathbf{A}$ and $\mathbf{B}$ are updated, while $\mathbf{W}_0$ remains frozen. During the supervised fine-tuning stage, we optimize the same objective from Equation (5) with trainable parameters θ:

$$\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{i \in \mathcal{Y}} \log p_\theta(y_i \mid y_{<i}, \mathbf{c}), \tag{7}$$

where θ comprises the LoRA parameters. This confines the optimization to a minimal set of parameters, enabling efficient adaptation while preserving the model's general reasoning capabilities.

4 Experiment

In this section, we first introduce the benchmarks, baselines, and experimental configurations used in our study. We then present the main detection and explanation results across four benchmarks. In addition, we conduct ablation studies to validate the encoder-free design and assess robustness to image degradation.

4.1 Experimental Setup

4.1.1 Benchmarks.

To comprehensively evaluate our model's performance in forgery detection and artifact explanation, we adopt four public datasets for training and evaluation. We leverage two datasets separately for training: FakeClue (Wen et al., 2025), a comprehensive dataset containing over 100K images across

Table 1: Detection and artifact explanation results on FakeClue and LOKI benchmarks. The best results are highlighted in boldface. † denotes models retrained on FakeClue.

Methods	FakeClue				LOKI			
	Acc ↑	F1 ↑	ROUGE_L ↑	CSS ↑	Acc ↑	F1 ↑	ROUGE_L ↑	CSS ↑
▼ *General-Purpose VLMs*								
Deepseek-VL2 (Wu et al., 2024)	47.5	54.1	17.2	50.5	43.1	39.2	16.9	38.8
InternVL2-40B (Chen et al., 2024c)	50.7	46.3	17.6	55.2	50.7	37.6	18.4	47.3
Qwen2-VL-72B (Wang et al., 2024)	57.8	56.5	17.5	54.4	55.4	40.9	17.3	43.2
GPT-4o (OpenAI, 2024)	47.4	42.0	13.4	40.7	63.4	57.2	14.7	35.4
NEO-9B (Diao et al., 2025)	52.4	48.8	17.3	53.9	51.6	38.2	16.4	39.3
▼ *Conventional Detectors (non-reasoning)*								
CNNSpot (Wang et al., 2020)	43.1	9.8	-	-	43.1	11.4	-	-
UnivFD (Ojha et al., 2023)	63.1	46.8	-	-	49.0	35.8	-	-
FreqNet (Tan et al., 2024b)	48.7	39.3	-	-	58.9	50.6	-	-
NPR (Tan et al., 2024a)	90.2	91.6	-	-	77.4	80.0	-	-
AIDE (Yan et al., 2025)	85.9	94.5	-	-	65.6	80.2	-	-
▼ *Forensic Experts*								
FakeVLM† (Wen et al., 2025)	96.5	95.8	57.4	**80.1**	82.6	81.9	18.4	55.5
AIGI-Holmes† (Zhou et al., 2025b)	96.2	95.4	53.9	72.0	80.8	78.3	18.7	51.2
Ivy-xDetector (Jiang et al., 2026)	92.4	94.1	47.2	59.7	76.7	74.9	16.4	43.5
EFG	**97.1**	**96.9**	**59.1**	79.0	**83.3**	**82.4**	**21.6**	**56.7**

seven categories with fine-grained artifact annotations; and Holmes-SFT Set (Zhou et al., 2025b), an instruction-tuning dataset comprising 65K images designed for out-of-distribution detection. For evaluation on held-out benchmarks, we follow prior work (Wen et al., 2025) and utilize LOKI (Ye et al., 2025), a benchmark with $2,217$ human-annotated test images, and DMImage (Corvi et al., 2023), a diffusion-based dataset of 21K test images widely used for comparison with conventional detectors.

4.1.2 Evaluation Metrics.

Our evaluation covers two dimensions: detection performance and artifact explanation quality. For detection, following established practices (Huang et al., 2025), we adopt accuracy (ACC) and macro-averaged F1 (Macro-F1) to assess classification correctness and robustness across generators. For explanation, we employ CSS and ROUGE-L to assess the generated rationales against existing annotated references, measuring semantic coherence and lexical fidelity respectively.

4.1.3 Baselines.

For a comprehensive evaluation of EFG, we conduct comparisons with baseline methods spanning three categories. **(1) General-purpose VLMs**, including GPT-4o (OpenAI, 2024), Qwen2-VL (Wang et al., 2024), DeepSeek-VL2 (Wu et al., 2024), InternVL2 (Chen et al., 2024c), and NEO (Diao et al., 2025), evaluated on detection benchmarks without further fine-tuning. **(2) Conventional detectors**, including CNNSpot (Wang et al., 2020), FreqNet (Tan et al., 2024b), UnivFD (Ojha et al., 2023), AIDE (Yan et al., 2025), and NPR (Tan et al., 2024a), which operate on low-level visual features without semantic reasoning. **(3) Fine-tuned forensic experts**, such as FakeVLM (Wen et al., 2025), AIGI-Holmes (Zhou et al., 2025b), and Ivy-xDetector (Jiang et al., 2026), representing state-of-the-art VLM-based detectors for synthetic image detection and explanation.

4.1.4 Training Settings.

Our model is initialized from the NEO-2.2B (Diao et al., 2025) checkpoint and undergoes parameter-efficient fine-tuning with LoRA on two training datasets separately. We train on FakeClue for 2 epochs with a learning rate of 1×10^{-3}, and on Holmes-SFT Set for 3 epochs with a learning rate of 2×10^{-5}, using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.95$ and weight decay 0.01. LoRA is applied to all linear projection matrices with rank $r = 128$ and scaling factor $\alpha = 256$. To reduce computational cost, the dynamic input resolution is constrained to the range from 256×256 to 1024×1024,

Table 2: Detection accuracy on Holmes-Set across various generator categories. J-1B and J-7B denote Janus-Pro-1B and Janus-Pro-7B, respectively; L-Gen denotes LlamaGen. The best results are highlighted in boldface, while the second-optimal results are marked with underline.

Methods	Generators										AVG (%)
	Janus	J-1B	J-7B	Show-o	L-Gen	Infinity	VAR	PixArt	SD3.5	FLUX	
▼ *Conventional Detectors*											
CNNSpot	70.0	70.9	85.0	72.2	61.9	86.8	59.9	78.2	63.8	79.9	72.9
UnivFD	87.6	96.9	96.4	85.9	93.1	79.2	64.3	75.7	87.8	69.6	83.6
FreqNet	64.2	68.4	89.1	75.3	84.9	87.4	66.0	72.6	88.5	73.2	76.9
NPR	51.2	69.5	73.9	93.7	93.5	93.8	85.9	93.4	91.6	93.6	84.0
AIDE	**91.2**	**98.9**	**97.8**	98.0	**99.4**	98.7	**93.6**	98.6	99.4	94.4	**97.0**
▼ *Forensic Experts*											
FakeVLM	76.8	89.7	90.3	93.9	95.0	98.2	65.8	98.9	99.2	98.7	90.6
AIGI-Holmes	80.2	91.9	89.6	**98.0**	98.0	98.4	76.0	98.5	97.8	94.2	92.3
Ivy-xDetector	62.8	84.6	87.6	88.6	97.5	99.3	66.1	92.8	96.9	93.6	87.0
EFG	81.6	95.2	96.4	91.6	93.1	**99.8**	67.4	**99.8**	**99.8**	**99.8**	92.5

and images outside this range will be resized to fit within it. The training is conducted with a per-device batch size of 16, gradient accumulation steps of 2, using a 3% linear warmup and a cosine decay schedule. We use bfloat16 mixed precision for efficient training. The proposed method is implemented in PyTorch, and all experiments are conducted on 4 A100 (40 GB) GPUs.

4.2 Main Results

We first evaluate our EFG on FakeClue (Wen et al., 2025), assessing both detection and explanation capabilities against three categories of baselines: general-purpose VLMs, conventional detectors, and fine-tuned forensic experts. Following the evaluation protocol of FakeVLM, we train EFG on FakeClue and evaluate it on both the FakeClue test set and LOKI (Ye et al., 2025), a cross-domain benchmark featuring human annotations. For fair comparison, all specialized forensic VLMs are trained on FakeClue following their original setups. Table 1 reports the detection and explanation performance of all methods.

As shown in the table, EFG delivers state-of-the-art performance, substantially outperforming both conventional detectors and specialized forensic VLMs. For instance, EFG surpasses the best conventional detector by 6.9% in detection accuracy on FakeClue. Our model also outperforms forensic experts, demonstrating its superior forensic detection capability. In terms of explanation quality, EFG also outperforms FakeVLM by 1.7% in lexical fidelity (ROUGE-L), highlighting its ability to generate more faithful and information-preserving descriptions. On the cross-domain LOKI benchmark with human annotations, EFG also achieves strong generalization, outperforming all specialized forensic VLMs by a notable margin. This is largely attributed to our native visual embedding design, which jointly preserves low-level signals while adapting to forensic reasoning, enabling the LLM backbone to access richer artifact cues and produce more informative, human-aligned explanations. In contrast, general-purpose VLMs exhibit negligible forensic sensitivity, achieving near-random accuracy on both benchmarks and failing to produce meaningful explanations for their predictions. This contrast underscores that semantic understanding alone is insufficient for forensic reasoning and that low-level artifact signals are essential for effective synthetic image detection.

We also evaluate EFG on the Holmes-Set benchmark to assess its out-of-distribution detection performance. This benchmark comprises images from ten categories generated by the latest unseen autoregressive visual generation models and diffusion models, providing a challenging testbed for generalization. As reported in Table 2, EFG demonstrates strong performance on this benchmark, surpassing specialized forensic VLMs including AIGI-Holmes. This confirms the effectiveness of our encoder-free visual embedding design, which preserves low-level artifact signals that enable EFG to reason over forensic cues in a way that other VLM-based detectors fail to exploit. We notice that Ivy-xDetector lags behind EFG by 5.5% in detection accuracy, and FakeVLM also underperforms EFG by a notable margin. This gap is largely due to their frozen visual encoders, which suppress fine-grained forensic traces during feature extraction.

Table 3: Comparison of detection methods on the full DMImage test set. Per-class accuracy and F1 are reported, with Overall F1 as the macro-average. $*$ denotes methods trained on FakeClue.

Methods	Real		Fake		Overall	
	Acc	F1	Acc	F1	Acc	F1
CNNSpot (Wang et al., 2020)	87.8	42.1	28.4	43.0	42.5	42.5
Fusing (Ju et al., 2022)	87.7	68.6	78.8	86.3	80.9	77.5
LNP (Bi et al., 2023b)	63.1	41.9	56.9	67.6	58.4	54.7
UnivFD (Ojha et al., 2023)	89.4	48.9	44.9	60.6	55.5	54.7
SIDA (Huang et al., 2025)	92.9	83.4	90.7	94.0	91.2	88.7
AIGI-Holmes* (Zhou et al., 2025b)	97.8	83.5	88.6	93.6	90.8	88.6
Ivy-xDetector* (Jiang et al., 2026)	94.3	84.2	90.7	94.2	91.6	89.2
FakeVLM* (Wen et al., 2025)	98.2	85.0	89.7	94.3	91.7	89.6
EFG*	**99.6**	**87.6**	**91.3**	**95.4**	**93.3**	**91.5**

However, we observe that EFG still underperforms the top-performing conventional detector, AIDE, on Holmes-Set. A closer breakdown reveals that although EFG performs well on diffusion-based categories, it struggles on autoregressive models, such as Janus (Chen et al., 2025b), where the model achieves notably lower accuracy. We attribute this discrepancy to the absence of autoregressive artifacts in the training data, which exposes a fundamental limitation of VLM-based forensic detectors, as these models lack specialized modules designed for cross-distribution generalization. This limitation highlights the inherent difficulty of out-of-distribution detection, where the rapid evolution of generative models continuously produces unseen forensic traces that challenge even well-designed forensic detectors.

In addition, we provide qualitative examples in Figure 3, comparing EFG with Qwen3-VL-Plus (Bai et al., 2025) and FakeVLM on the explanation task. EFG accurately identifies synthesis artifacts while generating coherent and precise natural language justifications. In contrast, the competing methods frequently yield incorrect predictions or generate explanations that are inconsistent with the visual evidence. For example, Qwen3-VL-Plus occasionally misclassifies synthetic images as authentic, while FakeVLM, despite reasonable detection, tends to produce generic or incomplete explanations that fail to pinpoint specific artifact regions. These comparisons underscore EFG's advantage in bridging low-level forensic evidence with semantic-level reasoning, producing faithful and coherent explanations grounded in visuals.

4.3 Cross-Dataset Generalization

To assess the generalization capability of EFG across distribution shifts, we conduct a cross-dataset evaluation by fine-tuning VLMs solely on FakeClue and directly testing on DMImage (Corvi et al., 2023). This benchmark allows us to compare against a broader range of detectors, including Fusing (Ju et al., 2022), LNP (Bi et al., 2023b), and SIDA (Huang et al., 2025). Table 3 reports the detection performance of EFG and competing methods on DMImage. EFG achieves competitive performance against state-of-the-art detectors, outperforming both conventional detectors and VLM-based methods. This result confirms that EFG generalizes well to DMImage despite having no exposure to it during training, suggesting that the model captures forensic signals that are not confined to specific training patterns within the diffusion domain. Notably, EFG also surpasses FakeVLM, the strongest competing forensic VLM, by 1.6% in detection accuracy, underscoring the advantage of our encoder-free design in cross-dataset generalization. This demonstrates the practical value of native VLMs in synthetic image detection, offering a readily deployable solution.

4.4 Ablation Study

4.4.1 Comparison with Encoder-Based VLMs

To isolate the effect of the encoder-free design, we compare EFG against three similarly sized encoder-based VLMs, Qwen2-VL-2B (Wang et al., 2024), InternVL3-2B (Zhu et al., 2025) and SmolVLM2-2.2B (Marafioti et al., 2025), each using a different visual encoder. We fine-tune all these models with the same LoRA configuration and a learning rate of 5×10^{-5}. The visual en-

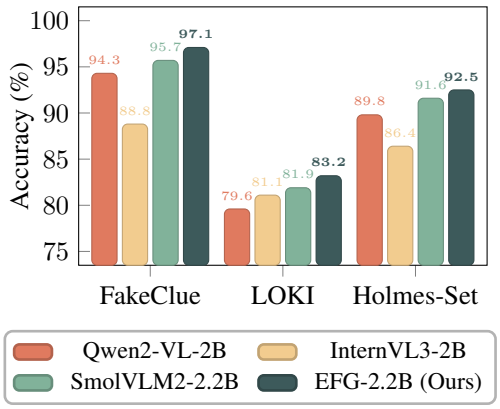

Figure 4: Performance evaluation of 2B-scale models with different visual encoders on three benchmarks.

Table 4: Robustness to JPEG compression and Gaussian blur on FakeClue. Classification accuracy is reported, with the best results highlighted in boldface.

Method	JPEG Compress		Gaussian Blur	
	QF=80	QF=70	$\sigma = 1$	$\sigma = 2$
CNNSpot	48.9	49.2	47.6	47.5
UnivFD	56.1	54.0	53.3	51.5
NPR	60.8	58.7	59.3	57.0
AIDE	70.4	66.5	68.7	65.3
FakeVLM	92.4	91.1	90.6	89.0
AIGI-Holmes	94.6	93.4	91.3	89.8
Ivy-xDetector	89.1	88.3	83.5	81.7
EFG	**95.5**	**94.7**	**92.0**	**90.9**

coders of these models are kept frozen during fine-tuning. As shown in Figure 4, EFG consistently outperforms its encoder-based counterparts across all three benchmarks, achieving substantial improvements in synthetic detection accuracy. The encoder-based VLMs also achieve reasonable detection performance across all benchmarks, indicating that certain artifact patterns are closely coupled with semantic features and remain partially extractable even after semantic compression. Among them, the detection performance varies considerably across benchmarks, suggesting that the choice of visual encoder heavily influences detection in a task-dependent manner. In particular, InternVL3-2B exhibits a notable performance gap on FakeClue and Holmes-Set. This is likely due to its visual encoder being distilled from a large model, a process that may compromise the feature extraction capability required for forensic detection.

4.4.2 Robustness to Degradation

We evaluate the robustness of EFG against two common types of image degradation: JPEG compression and Gaussian blurring. These distortions are known to alter low-level image statistics, which often degrade the performance of forensic detectors that rely on such fine-grained signals. For each type of perturbation, we evaluate two severity levels and summarize the results in Table 4. The results show that EFG exhibits strong resilience to both types of corruption across all tested settings. Notably, while VLM-based methods are relatively robust to these distortions, conventional detectors suffer from substantial performance degradation and tend to classify heavily compressed images as real, revealing a fundamental limitation of traditional forensic approaches. By leveraging both low-level and high-level semantic information, EFG mitigates the adverse effects of low-level distortions while also avoiding the secondary compression of already degraded fake clues through a pre-trained visual encoder. These results not only underscore the practical advantage of EFG in real-world scenarios but also position it as a more reliable foundation for forensic reasoning.

5 Conclusion

In this paper, we challenge the prevailing practice of using pre-trained visual encoders in vision-language models (VLMs) for AI-generated image detection. To address the bottleneck of low-level artifact suppression, we introduce EFG (Encoder-Free Forensic Guard), a native VLM built on unified primitives that reason across modalities, eliminating the need for pre-trained visual encoders. Our model preserves complete forensic clues and enables joint reasoning over low-level artifacts and high-level semantics, surpassing prior approaches suffering from encoder-induced trace loss, yielding accurate detection and grounded explanations. Extensive experiments demonstrate the effectiveness of EFG, which outperforms both encoder-based VLM forensic systems and conventional detectors in detection, while providing more explicit explanations. We believe our work will inspire the design of specialized VLMs for security-critical tasks, emphasizing native, artifact-preserving cross-modal reasoning.

AI USE STATEMENT

The authors confirm that large language models (LLMs) were used exclusively for polishing sentence structures and correcting grammatical errors. The LLMs were not involved in generating ideas, data, analyses, or any substantive content. The authors take full responsibility for the final content of this work, including any text revised with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All datasets used in this work are publicly available. Detailed training settings, including model configurations, hyperparameters, and optimization details, are described in Section 4.1.4. The source code is provided in the supplementary materials. The trained model weights will be released as open source upon acceptance. We believe that these resources, together with the details in the main text and supplementary materials, are sufficient to reproduce the reported results.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models. `https://www.adept.ai/blog/fuyu-8b`, 2023. Accessed: 2026-05-28.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey A. Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bosnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier J. Hénaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b VLM for transfer, 2024.

Xiuli Bi, Bo Liu, Fan Yang, Bin Xiao, Weisheng Li, Gao Huang, and Pamela C. Cosman. Detecting generated images by real images only, 2023a.

Xiuli Bi, Bo Liu, Fan Yang, Bin Xiao, Weisheng Li, Gao Huang, and Pamela C. Cosman. Detecting generated images by real images only, 2023b.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *ICML 2024*. OpenReview.net, 2024a.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes ai-generated image detector easier generalizable. In *NeurIPS 2025*, 2025a.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025b.

Yangyi Chen, Xingyao Wang, Hao Peng, and Heng Ji. A single transformer for scalable vision-language modeling. *Trans. Mach. Learn. Res.*, 2024, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR 2024*, pp. 24185–24198, 2024c.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023*, pp. 1–5. IEEE, 2023.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. Unveiling encoder-free vision-language models. In *NeurIPS 2024*, 2024.

Haiwen Diao, Mingxuan Li, Silei Wu, Linjun Dai, Xiaogang Wang, Hanming Deng, Lewei Lu, Dahua Lin, and Ziwei Liu. From pixels to words - towards native vision-language primitives at scale, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR 2021*. OpenReview.net, 2021.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS 2014*, pp. 2672–2680, 2014.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *CVPR 2025*, pp. 18685–18694. Computer Vision Foundation / IEEE, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS 2020*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR 2022*. OpenReview.net, 2022.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: social media image deepfake detection, localization and explanation with large multimodal model. In *CVPR 2025*, pp. 28831–28841. Computer Vision Foundation / IEEE, 2025.

Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can chatgpt detect deepfakes? A study of using multimodal large language models for media forensics. In *CVPR 2024*, pp. 4324–4333. IEEE, 2024.

Changjiang Jiang, Wenhui Dong, Zhonghao Zhang, Fengchang Yu, Wei Peng, Xinbin Yuan, Yifei Bi, Ming Zhao, Zian Zhou, Chenyang Si, and Caifeng Shan. Ivy-fake: A unified explainable framework and benchmark for image and video AIGC detection. In *ICMR 2026*, 2026.

Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *ICIP 2022*, pp. 3465–3469. IEEE, 2022.

Weixian Lei, Jiacong Wang, Haochen Wang, Xiangtai Li, Jun Hao Liew, Jiashi Feng, and Zilong Huang. The scalability of simplicity: Empirical analysis of vision-language learning with a single transformer. In *ICCV 2025*, pp. 20758–20769. IEEE, 2025.

Yixuan Li, Xuelin Liu, Xiaoyang Wang, Shiqi Wang, and Weisi Lin. Fakebench: Uncover the achilles' heels of fake images with large multimodal models, 2024.

Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Ke-Yue Zhang, Yue Zhou, Caiyong Piao, Bin Li, Taiping Yao, Bo Wang, Youchang Xiao, and Shouhong Ding. Deep residual injection for full-spectrum forensic signal perception in multimodal large language models, 2026.

Fengyuan Liu, Haochen Luo, Yiming Li, Philip Torr, and Jindong Gu. Which model generated this image? A model-agnostic approach for origin attribution. In *ECCV 2024*, volume 15120 of *Lecture Notes in Computer Science*, pp. 282–301. Springer, 2024.

Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiawen Liu, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. In *CVPR 2025*, pp. 24960–24971. Computer Vision Foundation / IEEE, 2025.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. Smolvlm: Redefining small and efficient multimodal models, 2025.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR 2023*, pp. 24480–24489. IEEE, 2023.

OpenAI. Gpt-4o. `https://openai.com/index/hello-gpt-4o`, 2024. Accessed: 2026-06-13.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Chuangchuang Tan, Huan Liu, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *CVPR 2024*, pp. 28130–28139. IEEE, 2024a.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *AAAI 2024*, pp. 5052–5060, 2024b.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2P-CLIP: injecting category common prompt in CLIP to enhance generalization in deepfake detection. In *AAAI 2025*, pp. 7184–7192. AAAI Press, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR 2020*, pp. 8692–8701. Computer Vision Foundation / IEEE, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *ICCV 2023*, pp. 22388–22398. IEEE, 2023.

Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation, 2025.

Shiyu Wu, Jing Liu, Jing Li, and Yequan Wang. Few-shot learner generalizes across ai-generated image detection. In *ICML 2025*. PMLR / OpenReview.net, 2025.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024.

Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In *ICLR 2025*. OpenReview.net, 2025.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *ICLR 2025*. OpenReview.net, 2025.

Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, Zhizheng Wu, Yiping Chen, Dahua Lin, Conghui He, and Weijia Li. LOKI: A comprehensive synthetic data detection benchmark using large multimodal models. In *ICLR 2025*. OpenReview.net, 2025.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. In *NeurIPS 2019*, pp. 12360–12371, 2019.

Yanran Zhang, Wenzhao Zheng, Yifei Li, Bingyao Yu, Yu Zheng, Lei Chen, Jiwen Lu, and Jie Zhou. Unigendet: A unified generative-discriminative framework for co-evolutionary image generation and generated image detection, 2026.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection, 2024.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *ICLR 2025*, 2025a.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models. In *ICCV 2025*, pp. 18746–18758. IEEE, 2025b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In *NeurIPS 2023*, 2023.