FORENSICZOOM: ADAPTIVE VISUAL INSPECTION WITH MULTIMODAL LLMS FOR INDUSTRIAL-GRADE FACE FORGERY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable face forgery detection is critical to the security of online identity verification systems, where missed attacks compromise security and excessive false positives disrupt legitimate users. Specialized forensic detectors achieve strong detection performance but provide limited interpretability, while multimodal large language models (MLLMs) offer strong semantic understanding and interpretable reasoning yet remain substantially weaker for face forgery detection. We argue that a key limitation lies in how visual evidence is acquired: subtle forensic artifacts may be poorly represented at standard resolution, while uniformly processing all cases at higher resolution is computationally inefficient. We therefore introduce **ForensicZoom**, an industrial-grade MLLM framework for *adaptive visual inspection*. ForensicZoom first equips a general-purpose MLLM with forensic-aware visual representations and aligns the language model with these features. Its central mechanism, `NEED_ZOOM`, enables the model to autonomously request magnified views of suspicious regions when the initial evidence is insufficient, turning fixed-pass classification into adaptive multi-round forensic reasoning. The zoom behavior is learned through reward shaping that balances detection accuracy with unnecessary visual inspection, concentrating additional computation on difficult cases. A final attribution optimization stage improves natural-language forensic reports while preserving detection performance. On large-scale industrial identity verification data, ForensicZoom achieves over 97% TPR at 0.1% FPR, substantially outperforming both specialized detectors and existing MLLM-based methods while producing actionable forensic attributions. These results demonstrate that ForensicZoom can provide an effective path toward accurate, interpretable, and scalable MLLM-based face forgery detection. Our scripts and model parameters will be released upon acceptance.

1 INTRODUCTION

As generative AI becomes increasingly capable and widely deployed, concerns over the safety and misuse of AI-generated content continue to grow Goodfellow et al. (2014); Ho et al. (2020); Rombach et al. (2022). Face forgery is a particularly consequential instance of this problem: malicious actors can manipulate or synthesize facial imagery to impersonate legitimate users and bypass online identity verification systems, creating substantial risks for financial services, e-government platforms, and other security-sensitive applications Sumsub (2024); Jia et al. (2024); Chandra et al. (2025). Reliable face forgery detection is therefore essential, especially in industrial settings where missed attacks directly compromise security while excessive false positives disrupt legitimate users. Beyond accuracy, interpretability is important for supporting human review of ambiguous cases with clear forensic evidence. This challenge is becoming increasingly difficult as modern generative models produce highly realistic faces whose forensic traces may be limited to subtle local artifacts, further obscured by compression, heterogeneous capture conditions, and rapidly evolving generation pipelines Chandra et al. (2025); Jiang et al. (2020).

To address this problem, existing approaches primarily rely on specialized forensic models or multimodal large language models (MLLMs). Specialized CNN- and ViT-based detectors, including XceptionNet Chollet (2017), EfficientNet Tan & Le (2019), SBI Shiohara & Yamasaki (2022), and

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.

LAA-Net Nguyen et al. (2024), achieve strong detection performance but offer limited interpretability, typically producing predictions without explicit forensic reasoning or natural-language attribution. In contrast, MLLMs can jointly process visual and contextual information while providing interpretable explanations, but general-purpose MLLMs remain substantially weaker than specialized detectors for face forgery detection Jia et al. (2024); Xu et al. (2025). Recent methods improve MLLM-based forensics through task-specific fine-tuning, expert-assisted prediction, artifact-aware supervision, reinforcement learning, and external forensic tools Huang et al. (2024); Wen et al. (2025); Liu et al. (2024); Xu et al. (2026); Zhou et al. (2026); Wang et al. (2026). Nevertheless, current MLLM-based methods still struggle to capture the fine-grained forensic evidence required for reliable detection and remain far from the stringent accuracy and efficiency requirements of industrial deployment.

Our experience with large-scale identity verification suggests that a central limitation lies in how visual evidence is acquired and allocated during inference. Human vision is known to follow a coarse-to-fine process, forming an initial global perception before allocating focused attention to local details when necessary Navon (1977); Treisman & Gelade (1980); Gerlach & Poirel (2020). Recent MLLM research has begun to adopt a similar principle through guided visual search and active perception, allowing models to selectively inspect task-relevant regions rather than processing all visual information uniformly Wu & Xie (2023); Zhu et al. (2026). Face forgery detection naturally exhibits this structure: many cases can be resolved from a global observation, while difficult cases may depend on subtle local artifacts that fall below the effective resolution of standard visual encoding. Applying high-resolution analysis uniformly is computationally inefficient at production scale, whereas fixed-resolution single-pass inference may overlook critical forensic evidence. These observations motivate an *adaptive visual inspection* paradigm, where an MLLM selectively acquires higher-resolution evidence from suspicious regions when needed.

A key limitation of existing face forgery benchmarks is that they often evaluate models on isolated media inputs under distributions that do not fully reflect real-world deployment conditions Chandra et al. (2026); Li et al. (2026). Recent studies have increasingly emphasized multimodal evaluation, incorporating complementary visual, acoustic, depth, or infrared signals to better approximate practical face-security settings Yu et al. (2023); Shi et al. (2025). In practice, however, behavioral information associated with the acquisition process—such as the time spent capturing an image, repeated verification attempts, and other interaction signals—can provide critical complementary evidence for distinguishing sophisticated attacks from legitimate users, particularly at the extremely low false-positive rates required in production. The absence of such signals creates a substantial gap between benchmark evaluation and industrial deployment: a pipeline designed and optimized solely for existing visual benchmarks may not directly translate to a practical verification system where visual and behavioral evidence must be jointly considered. To study face forgery detection under this more realistic setting, we therefore evaluate on a private multimodal dataset collected from an operational online identity verification system serving thousands of requests per day, which contains both visual inputs and associated behavioral information.

In this work, we introduce **ForensicZoom**, an industrial-grade MLLM framework for adaptive face forgery detection. We adopt an MLLM rather than a conventional vision-only detector because practical identity verification requires not only accurate predictions but also interpretable forensic evidence that can support subsequent human review (Rudin, 2019; Guo et al., 2025; Yu et al., 2025). ForensicZoom first equips a general-purpose MLLM with forensic-aware visual representations and aligns the language model with these features. We further observe a substantial performance gap between global and locally magnified views for subtle forgery cases, indicating that critical forensic cues can be lost under fixed-resolution visual encoding (Zhao et al., 2021; Liu et al., 2021a). Its central mechanism, `NEED_ZOOM`, therefore enables the model to autonomously request magnified views of suspicious regions when the initial evidence is insufficient, turning fixed-pass classification into adaptive multi-round forensic reasoning. The zoom behavior is learned through reward shaping that balances detection accuracy with unnecessary visual inspection, concentrating additional computation on difficult cases. A final attribution optimization stage improves natural-language forensic reports while preserving detection performance. On large-scale industrial identity verification data, ForensicZoom achieves over 97% TPR at 0.1% FPR while producing actionable forensic attributions, and has been deployed in a production pipeline processing millions of requests per day.

We summarize our contributions as follows:

- We introduce NEED_ZOOM, a learned mechanism that enables an MLLM to recognize insufficient forensic evidence and autonomously request higher-resolution inspection of suspicious regions.
- We develop **ForensicZoom**, a progressive training framework that builds forensic perception, aligns it with language reasoning, and enables adaptive multi-round inference with interpretable attribution.
- We validate ForensicZoom on large-scale industrial identity verification data, achieving over 97% TPR at 0.1% FPR, and demonstrate its effectiveness through production deployment at millions of requests per day.

2 Related Work

2.1 Face Forgery Detection

Face forgery refers to manipulated or generated facial imagery that alters a person's identity, appearance, or authenticity. Early detection methods exploit low-level forensic cues such as inconsistent head poses (Yang et al., 2019), abnormal eye blinking (Li et al., 2018), and frequency artifacts (Qian et al., 2020). Deep learning approaches learn forensic representations directly using architectures such as XceptionNet (Chollet, 2017), EfficientNet (Tan & Le, 2019), and Vision Transformers (Coccomini et al., 2022). SBI (Shiohara & Yamasaki, 2022) improves generalization through pseudo-forgery synthesis. LAA-Net (Nguyen et al., 2024) focuses on localized manipulation artifacts. FSFM (Wang et al., 2025) learns self-supervised facial representations. Frequency-based methods explicitly model spectral inconsistencies (Qian et al., 2020; Liu et al., 2021b). More recently, MLLMs have been explored for interpretable forgery detection. DD-VQA (Zhang et al., 2024) formulates detection as visual question answering, FFAA (Huang et al., 2024) incorporates an external expert classifier, FakeVLM (Wen et al., 2025) fine-tunes on artifact annotations, and ForgeryGPT (Liu et al., 2024) introduces identity-aware representations. MARE (Xu et al., 2026) applies RLHF for attribution alignment. EvolveReason (Zhou et al., 2026) uses GRPO for self-evolving forensic reasoning. Jiang et al. (2026) combines reinforcement learning with Self-Blended Images. ForgeryVCR (Wang et al., 2026) trains an MLLM to invoke external forensic tools.

2.2 AI Content Safety

The rapid development of generative models, from GANs (Goodfellow et al., 2014; Karras et al., 2019) to diffusion models (Ho et al., 2020; Rombach et al., 2022), has substantially improved the realism and diversity of AI-generated visual content, creating growing challenges for content authenticity and safety. Early detection methods identify synthetic images through low-level statistical artifacts, including color inconsistencies (McCloskey & Albright, 2018) and generator-specific fingerprints (Yu et al., 2019). Subsequent work investigates generalization across unseen generators (Wang et al., 2020; Ojha et al., 2024), while GenImage (Zhu et al., 2023) provides a large-scale benchmark for evaluating AI-generated image detection across diverse generative models. Diffusion-specific approaches further exploit reconstruction characteristics of diffusion models, such as DIRE (Wang et al., 2023) and AEROBLADE (Ricker et al., 2024). Beyond post-hoc detection, watermarking methods such as Tree-Ring Watermarks (Wen et al., 2023) embed identifiable signals directly into generated images to facilitate provenance verification. Mechanistic interpretability modules have been utilized to analyze visual patterns associated with generative-model failures: including Language-Grounded Sparse Encoders for general content analysis (Tang et al., 2026), and Matryoshka Transcoders for physical plausibility failure modes (Tang et al., 2025).

3 Method

In this section, we introduce the methodology of ForensicZoom (Figure 1). Section 3.1 first enhances the vision encoder with forensic perception, and Section 3.2 aligns the language model with these forensic visual signals through supervised fine-tuning. Section 3.3 then introduces the NEED_ZOOM mechanism, which enables adaptive higher-resolution inspection. Finally, Section 3.4 presents Cascade RL for improving forensic attribution while preserving detection performance.

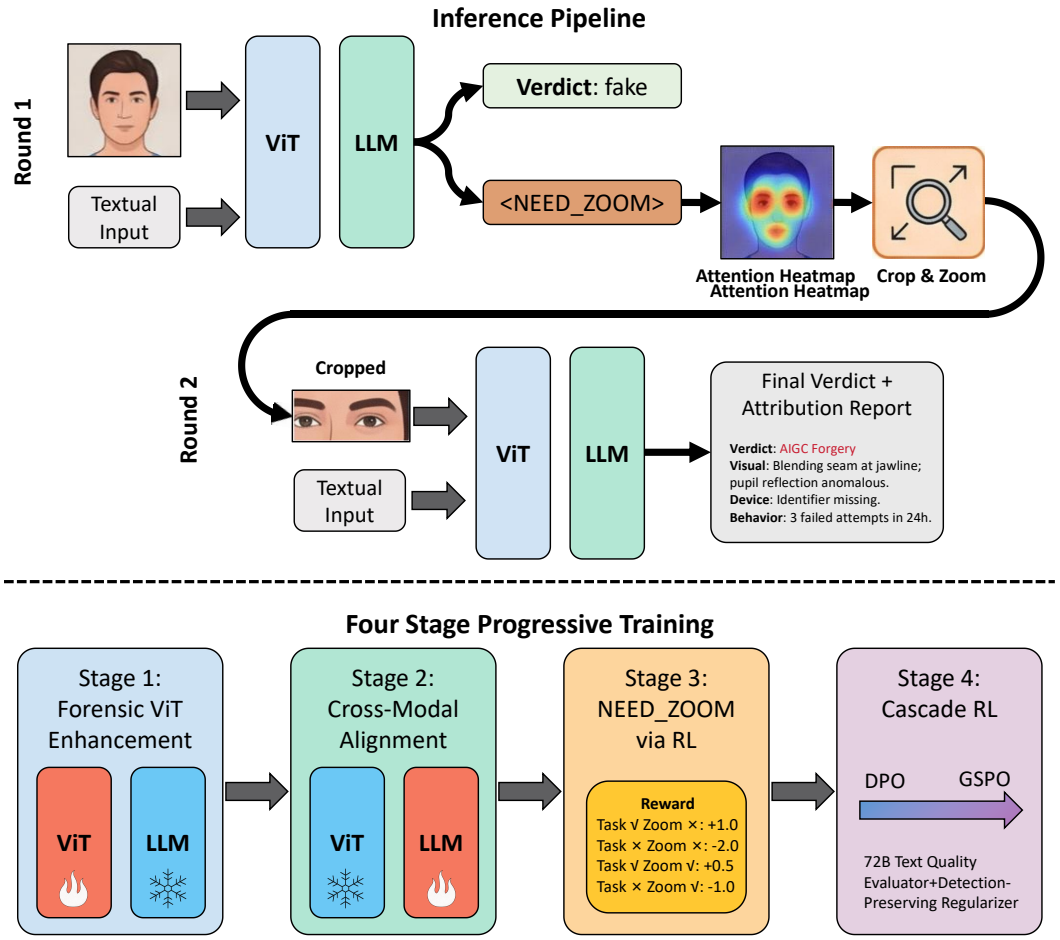

Figure 1: Overview of ForensicZoom. **Top:** Inference pipeline. In Round 1, the face image and text prompt are processed by the ViT (with Forensic Query Bank) and LLM. If the model outputs <NEED_ZOOM>, the attention heatmap guides cropping and magnification of suspicious regions. In Round 2, the cropped patches are fed back through the same pipeline to produce the final verdict. **Bottom:** Four-stage progressive training. Stage 1 enhances the ViT with forensic perception. Stage 2 aligns the LLM with enhanced visual signals. Stage 3 trains the <NEED_ZOOM> decision policy via PPO with asymmetric rewards. Stage 4 optimizes attribution quality through Cascade RL, transitioning from DPO to GSPO with a text quality evaluator and detection-preserving regularizer.

3.1 Forensic Vision Enhancement

We first adapt the vision encoder to capture the fine-grained, high-frequency cues that distinguish genuine faces from forgeries, and compress them into a compact set of forensic features. In specific, we freeze the LLM $\mathcal{L}$ and the projection layer $\mathcal{P}$, and unfreeze the last K layers of the ViT encoder $\mathcal{V}$, approximately 25–30% of total ViT layers. Let $\mathcal{V} = [\mathcal{V}^{(1)}, \ldots, \mathcal{V}^{(L)}]$ denote the L-layer ViT, where layers 1 through $L - K$ remain frozen and layers $L - K + 1$ through L are trainable. Lower layers capture general visual features such as edges and textures that transfer well, while upper layers encode task-specific representations that require adaptation.

Forensic Query Bank (FQB). We introduce a set of M learnable query tokens $\mathbf{Q} = \{\mathbf{q}_1, \ldots, \mathbf{q}_M\} \in \mathbb{R}^{M \times d}$, where d is the ViT hidden dimension, trained to extract forensic-relevant information from the ViT output through a cross-attention mechanism.

Let $\mathbf{Z} = \mathcal{V}(\mathbf{I}) \in \mathbb{R}^{N \times d}$ denote the N visual tokens output by the ViT. The FQB applies multi-head cross-attention:

$$\mathbf{F} = \text{CrossAttn}(\mathbf{Q}, \mathbf{Z}, \mathbf{Z}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{W}_Q(\mathbf{Z}\mathbf{W}_K)^\top}{\sqrt{d_k}}\right)\mathbf{Z}\mathbf{W}_V, \tag{1}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d_k}$ are learnable projection matrices and $\mathbf{F} \in \mathbb{R}^{M \times d}$ are the resulting *forensic tokens*. These tokens "query" and disentangle the high-frequency forensic features from the entangled ViT representations, where semantic and forensic information are intertwined.

The forensic tokens are concatenated with the original visual tokens before being fed into the LLM:

$$\mathbf{H} = [\mathcal{P}(\mathbf{Z});\ \mathcal{P}_f(\mathbf{F})] \in \mathbb{R}^{(N+M) \times d_l}, \tag{2}$$

where $\mathcal{P}$ and $\mathcal{P}_f$ are MLP projection layers mapping visual and forensic tokens respectively into the LLM embedding space of dimension d_l.

Attention Entropy Guidance. To encourage focused rather than diffuse attention, we add an attention entropy regularization term. For each attention head h in the trainable ViT layers, let $\mathbf{A}^{(h)} \in \mathbb{R}^{N \times N}$ denote the attention weight matrix. We compute the per-token entropy:

$$\mathcal{H}_i^{(h)} = -\sum_{j=1}^{N} A_{ij}^{(h)} \log A_{ij}^{(h)}, \tag{3}$$

and define the entropy regularization loss as:

$$\mathcal{L}_{\text{ent}} = \frac{1}{|\mathcal{H}_{\text{heads}}|} \sum_{h \in \mathcal{H}_{\text{heads}}} \frac{1}{N} \sum_{i=1}^{N} \mathcal{H}_i^{(h)}, \tag{4}$$

where $\mathcal{H}_{\text{heads}}$ is the set of attention heads in trainable layers. Minimizing $\mathcal{L}_{\text{ent}}$ concentrates attention on the most discriminative regions. The total loss for this stage combines the standard cross-entropy loss for binary classification with the entropy regularization is $\mathcal{L}_{\text{S1}} = \mathcal{L}_{\text{CE}}(y, \hat{y}) + \lambda_{\text{ent}} \cdot \mathcal{L}_{\text{ent}}$, where $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss on the detection token (real/fake), and λ_{ent} is a weighting coefficient.

3.2 Cross-Modal Alignment via SFT

We next align the language model with these forensic visual signals using standard supervised fine-tuning, enabling it to interpret and reason over the features learned in the previous stage.

Training Objective. The loss function for this stage combines three terms:

$$\mathcal{L}_{\text{S2}} = \mathcal{L}_{\text{CE}}(y, \hat{y}) + \lambda_{\text{mix}} \cdot \mathcal{L}_{\text{CE}}(\tilde{y}, \hat{\tilde{y}}) - \lambda_{\text{cal}} \cdot \mathcal{L}_{\text{cal}}, \tag{5}$$

where the second term is a feature-space mixup loss Zhang et al. (2018) that interpolates visual features of real/fake pairs ($\tilde{\mathbf{H}} = \alpha\mathbf{H}_i + (1-\alpha)\mathbf{H}_j$, $\alpha \sim \text{Beta}(0.4, 0.4)$) to strengthen the decision boundary, and the third term is an output entropy regularization $\mathcal{L}_{\text{cal}} = -\sum_v p(v) \log p(v)$ over the detection vocabulary, maximized to prevent overconfident predictions. Since the next stage relies on the model's implicit uncertainty to determine when additional evidence is needed, this calibration helps produce more reliable zoom decisions.

3.3 The NEED_ZOOM Mechanism

We next introduce NEED_ZOOM, the core mechanism that enables ForensicZoom to recognize insufficient visual evidence and selectively request higher-resolution inspection before making decisions.

Adaptive Zoom Inference. We extend the model vocabulary with a special token <NEED_ZOOM>. During inference, the model first examines the original image and either produces a verdict directly or emits <NEED_ZOOM> together with a textual specification of the suspicious region (e.g., "eye region shows suspicious symmetry"). When <NEED_ZOOM> is generated, an external tool uses the ViT's last-layer attention map to localize the most attended region, crops a fixed 48×48 window from the original 256×256 image, resizes it to 256×256, and constructs a second-round input containing both the original image and the magnified crop through Qwen3-VL's native multi-image

capability. The model then re-examines the visual evidence and produces the final verdict and attribution. Otherwise, the prediction is completed in a single round.

Importantly, the decision to emit <NEED_ZOOM> does not rely on an explicit confidence threshold. Instead, the model learns when additional visual evidence is necessary through reinforcement learning, allowing the zoom policy to emerge directly from the detection objective.

Reward Design. We optimize the NEED_ZOOM policy using Proximal Policy Optimization (PPO) Schulman et al. (2017). The reward jointly accounts for detection accuracy and the cost of additional inspection: a correct decision without zoom yields $+1.0$; a missed detection without zoom incurs -2.0; a zoom followed by a correct decision yields $+0.5$; and a zoom followed by an incorrect decision incurs -1.0. This asymmetric design encourages the model to request magnification when the initial evidence is insufficient, while assigning a lower reward to successful zooming than to a correct single-round prediction discourages unnecessary inspection.

CoT-Attention Alignment. To ensure that the requested region is consistent with the model's visual evidence, we align the spatial distribution implied by its textual region description with the ViT's last-layer attention map using $\mathcal{L}_{\text{align}} = D_{\text{KL}}(\mathbf{a}_{\text{text}} \| \mathbf{a}_{\text{vis}})$. This regularization discourages hallucinated region descriptions and grounds the zoom request in the model's visual attention. The final objective combines the standard PPO clipped surrogate Schulman et al. (2017) with the alignment loss:

$$\mathcal{L}_{\text{S3}} = \mathcal{L}_{\text{PPO}}(\theta; R) + \lambda_{\text{align}} \cdot \mathcal{L}_{\text{align}}. \tag{6}$$

3.4 CASCADE RL

We finally introduce *Cascade RL* to improve natural-language forensic attribution while preserving detection performance. Cascade RL consists of two consecutive optimization phases: we first apply standard Direct Preference Optimization (DPO) for stable preference alignment, and then transition to Group Sequence Policy Optimization (GSPO) Zheng et al. (2025) for online refinement. This design combines the stability of offline preference optimization with the stronger exploration and refinement capability of online reinforcement learning.

DPO Phase. We first optimize the standard DPO objective:

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(a_w, a_l)} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(a_w|x)}{\pi_{\text{ref}}(a_w|x)} - \beta \log \frac{\pi_\theta(a_l|x)}{\pi_{\text{ref}}(a_l|x)} \right) \right], \tag{7}$$

where $x = (\mathbf{I}, \mathbf{c})$, π_{ref} is the reference policy, and β is the temperature parameter.

GSPO Phase. We then apply GSPO, which defines a length-normalized sequence-level importance ratio $s_i(\theta) = \left(\frac{\pi_\theta(a_i|x)}{\pi_{\theta_{\text{old}}}(a_i|x)} \right)^{1/|a_i|}$ for each sampled response a_i, together with the group-relative advantage $\hat{A}_i = \frac{R_{\text{attr}}(a_i) - \bar{R}}{\sigma_R}$, where $\bar{R} = \frac{1}{G} \sum_{j=1}^{G} R_{\text{attr}}(a_j)$. The GSPO objective is

$$\mathcal{J}_{\text{GSPO}}(\theta) = \frac{1}{G} \sum_{i=1}^{G} \min \Big(s_i(\theta) \hat{A}_i,\ \text{clip}(s_i(\theta), 1-\epsilon, 1+\epsilon) \hat{A}_i \Big). \tag{8}$$

Cascade Transition. We gradually transition from DPO to GSPO over T optimization rounds:

$$\mathcal{L}_{\text{RL}}^{(t)} = (1 - \alpha_t) \cdot \mathcal{L}_{\text{DPO}} - \alpha_t \cdot \mathcal{J}_{\text{GSPO}}, \quad \alpha_t = \min \left(1, \frac{t}{T} \right), \tag{9}$$

where t denotes the current round. In practice, we use $T = 4$ rounds with $\alpha_t \in \{0, 0.2, 0.5, 1.0\}$.

Preference Data Construction. To provide supervision for attribution optimization, we employ a large language model as a text quality evaluator that scores clarity, coherence, and actionability. For each sample, we generate $G = 5$ attributions at varied temperatures, score each by $R_{\text{attr}}(a_i) = w_1 S_{\text{evidence}} + w_2 S_{\text{logic}} + w_3 S_{\text{action}}$, and construct preference pairs from the top- and bottom-ranked responses.

Detection-Preserving Regularization. Since optimizing attribution quality can degrade detection performance, we retain the detection objective throughout Cascade RL:

$$\mathcal{L}_{\text{S4}} = \mathcal{L}_{\text{RL}}^{(t)} + \lambda_{\text{det}} \cdot \mathcal{L}_{\text{CE}}(y, \hat{y}), \tag{10}$$

where λ_{det} controls the detection-preservation term. We use $\lambda_{\text{det}} = 0.1$ in our experiments.

4 EXPERIMENTS

In this section, we evaluate ForensicZoom on large-scale industrial identity verification data. Section 4.1 describes the dataset, implementation details, evaluation metrics, and baselines. Section 4.2 compares ForensicZoom with specialized detectors and MLLM-based methods on face forgery detection. Section 4.3 analyzes the contributions of the individual components, with particular attention to the `NEED_ZOOM` mechanism and Cascade RL. Section 4.4 evaluates the quality of the generated forensic attributions, and Section 4.5 provides qualitative examples of adaptive visual inspection and forensic reasoning.

4.1 SETUP

Dataset. All experiments use a large-scale industrial identity verification dataset from a production platform, spanning 8 months (April–December 2025). This period was chosen because diffusion-based synthesis became the dominant attack vector in production starting early 2025, largely displacing earlier GAN-based methods. Attacks fall into two categories: *AIGC-based forgeries* (diffusion-based face swapping, single-image animation) and *presentation attacks* (screen replay, printed photos). AIGC attacks are substantially harder to detect and represent the primary focus. The dataset statistics are summarized in Table 1. Note that the training set is curated with deliberate oversampling of attack samples; the real production distribution has a far higher legitimate-to-attack ratio (exceeding 1000:1), which is why FPR = 0.1% is the only operationally meaningful threshold. The dataset is split 80/10/10 for train/val/test with no user overlap. Due to confidentiality, the dataset and code are not publicly available.

Table 1: Dataset statistics.

Category	Count	Proportion
Legitimate (real)	2.0M	97.6%
AIGC forgery (swap + animation)	20K	1.0%
Presentation attack (replay + print)	30K	1.4%
Total	2.05M	100%

Implementation. We use Qwen3-VL-8B-Instruct Bai et al. (2025) as the base MLLM. The ViT has 27 transformer layers with hidden dimension $d = 1152$. In Stage 1, we unfreeze the last $K = 7$ layers and use $M = 32$ FQB query tokens. Training runs for 3 epochs at learning rate 5×10^{-5} (ViT) and 1×10^{-4} (FQB), with $\lambda_{\text{ent}} = 0.1$, batch size 32 on 8 GPUs (80GB) with DeepSpeed ZeRO-2. Stage 2 fine-tunes the LLM for 2 epochs at 2×10^{-5}, with $\lambda_{\text{mix}} = 0.5$, $\lambda_{\text{cal}} = 0.01$. Stage 3 runs 5000 PPO iterations at 1×10^{-6}, with $\epsilon = 0.2$, $\lambda_{\text{align}} = 0.1$. Stage 4 uses $\beta = 0.1$ for DPO, $\epsilon = 0.02$ for GSPO sequence-level clipping (significantly tighter than token-level methods due to the different numerical range of sequence-level importance ratios Zheng et al. (2025)), with $\lambda_{\text{det}} = 0.1$.

Metrics. We report TPR at FPR=0.1%, following industrial convention where the extreme class imbalance (legitimate-to-attack ratio exceeds 1000:1) makes stricter FPR thresholds the only operationally meaningful metric. We also report AUC. Attribution quality is evaluated by three expert reviewers on 200 sampled flagged cases, rating evidence completeness (S_{evidence}), logical coherence (S_{logic}), and actionability (S_{action}) on a 1–5 scale, averaged as $Q = \frac{1}{3}(S_{\text{evidence}} + S_{\text{logic}} + S_{\text{action}})$.

Baselines. We compare against: (1) *Specialized models*: XceptionNet Chollet (2017), EfficientNet-B4 Tan & Le (2019), SBI Shiohara & Yamasaki (2022), LAA-Net Nguyen et al. (2024); (2) *General VLMs* (zero-shot): GPT-4o, Qwen3-VL-8B; (3) *VLM-based methods* (re-trained on our data): FFAA Huang et al. (2024), FakeVLM Wen et al. (2025), MARE Xu et al. (2026).

4.2 MAIN RESULTS

Table 2 presents overall detection results. ForensicZoom achieves 97.3% TPR at FPR=0.1%, surpassing the best specialized model (LAA-Net, 91.0%) by 6.3 percentage points while simultane-

Table 2: Main detection results on AIGC forgeries. †: re-trained on our data.

Type	Method	TPR@0.1%	AUC	Attr.
Specialized Models	XceptionNet	89.2	.981	✗
	EfficientNet-B4	90.5	.984	✗
	SBI	88.3	.978	✗
	LAA-Net	91.0	.985	✗
General VLMs	GPT-4o	31.5	.742	✓
	Qwen3-VL	15.8	.661	✓
VLM-based Detection	FFAA†	82.6	.967	✓
	FakeVLM†	85.1	.972	✓
	MARE†	86.8	.975	✓
	ForensicZoom	**97.3**	**.998**	✓

Table 3: Per-attack-type TPR@FPR=0.1%.

Attack Type	Best Baseline	Ours	Zoom %
AIGC forgery	91.0	**97.3**	23%
Presentation attack	98.5	**99.8**	3%

ously providing natural-language attributions. General-purpose VLMs perform poorly in zero-shot settings (Qwen3-VL achieves only 15.8% TPR on our data), confirming the forensic perception deficit. Existing VLM-based methods improve over zero-shot but remain below specialized models, reflecting the limitations of indirect ViT optimization.

Table 3 breaks results down by attack category. On AIGC forgeries, ForensicZoom achieves 97.3% TPR compared to the best baseline of 91.0%, with a zoom trigger rate of 23%. On presentation attacks, which are inherently easier due to more prominent artifacts, all methods perform well and ForensicZoom reaches 99.8% with minimal zooming (3%).

4.3 Ablation Study

Stage-wise contribution. Table 4 shows the incremental contribution of each stage. Stage 1 lifts TPR from 15.8% to 88.5%, establishing forensic perception. Stage 2 further improves to 93.7% through LLM alignment. Stage 3 adds interactive zoom, reaching 96.9%. Stage 4 yields a further 0.4pp TPR gain while substantially improving attribution quality from 2.85 to 4.21. The modest TPR improvement from Stage 4 can be attributed to the Cascade RL training requiring the model to organize more rigorous reasoning chains when generating attributions, which indirectly disciplines its decision process and reduces borderline misclassifications.

Table 4: Stage-wise ablation.

Configuration	TPR@0.1%	Attr. Q	Zoom
Base (zero-shot)	15.8	–	–
+ Stage 1 (ViT+FQB)	88.5	–	–
+ Stage 2 (LLM SFT)	93.7	–	–
+ Stage 3 (PPO+Zoom)	96.9	2.85	18%
+ Stage 4 (Cascade RL)	**97.3**	**4.21**	17%

FQB and attention entropy. Removing FQB from Stage 1 drops TPR from 88.5% to 84.1%, and further removing attention entropy guidance drops it to 81.3%. Performance plateaus at $M = 32$ query tokens; increasing to $M = 64$ yields negligible improvement (+0.2pp).

Zoom behavior. All zoom trigger rates reported in this section are computed over the full test set. The model exhibits rational zoom behavior. On easy samples (first-round logit margin > 3.0), the

zoom rate is only 4%, while on hard samples (margin < 1.0) it reaches 47%. Zooming improves TPR on hard samples from 78.2% to 93.6%, a 15.4pp gain. The overall zoom rate of 17% means the majority of samples are resolved in a single round. The slight decrease from 18% (Stage 3) to 17% (Stage 4) reflects that attribution training refines the model's internal confidence calibration, marginally raising the implicit zoom threshold.

Cascade RL strategy. Table 5 compares Stage 4 strategies. DPO alone improves attribution quality but drops TPR by 3.8pp. GSPO alone causes a 5.2pp drop due to aggressive updates. Cascade RL achieves the best attribution score while *increasing* TPR by 0.4pp, validating the progressive transition. The detection-preserving regularizer is critical: removing it leads to a 2.6pp TPR drop.

Table 5: Stage 4 RL strategy ablation. TPR at FPR=0.1%.

Strategy	**TPR**	**Attr.** Q	Δ**TPR**
No Stage 4	96.9	2.85	–
DPO only	93.1	3.92	-3.8
GSPO only	91.7	3.78	-5.2
Cascade RL	**97.3**	**4.21**	$+0.4$
w/o det. reg.	94.3	4.15	-2.6

4.4 Attribution Quality

Table 6 presents human evaluation results (Fleiss' $\kappa = 0.71$). ForensicZoom achieves the highest scores across all dimensions. The largest gap over baselines appears in evidence completeness, reflecting the model's ability to incorporate fine-grained visual details obtained through zooming. Stage 4 provides a substantial boost, particularly in actionability ($+1.2$ over the w/o-S4 variant).

Table 6: Attribution quality (1–5 scale, 200 cases).

Method	S_{evi}	S_{log}	S_{act}	Q
GPT-4o (zero-shot)	2.1	2.8	1.9	2.27
Qwen3-VL (zero-shot)	1.7	2.3	1.5	1.83
FFAA†	2.8	3.1	2.5	2.80
MARE†	3.2	3.4	2.9	3.17
ForensicZoom (w/o S4)	3.5	3.3	3.0	3.27
ForensicZoom	**4.3**	**4.1**	**4.2**	**4.20**

4.5 Qualitative Analysis

Fig. 1 (top) illustrates the two-round inference. On easy cases, the model directly outputs a verdict. On hard cases, it generates observational text describing ambiguous regions, emits `<NEED_ZOOM>`, and after receiving magnified crops, produces a detailed attribution citing specific artifacts such as blending seams at the jawline and symmetric pupil reflections inconsistent with natural capture. ViT attention maps after Stage 1 concentrate sharply on forged boundaries and texture discontinuities, whereas the pre-trained ViT distributes attention broadly across the face.

5 Conclusion

We presented **ForensicZoom**, an industrial-grade MLLM framework for adaptive face forgery detection. By enhancing forensic perception, aligning visual evidence with language reasoning, and introducing the learned `NEED_ZOOM` mechanism, ForensicZoom selectively acquires higher-resolution evidence only when necessary while producing interpretable forensic attributions. Experiments on large-scale industrial identity verification data demonstrate strong detection performance, achieving 97.3% TPR at 0.1% FPR and substantially outperforming both specialized detectors and existing MLLM-based methods. These results suggest that adaptive visual inspection provides a promising path toward accurate, interpretable, and scalable face forgery detection.

AI USE STATEMENT

In this work, we used generative AI tools, including ChatGPT and Claude to implement methods, write and edit code, and assist in drafting the manuscript. We did not use generative AI tools to generate synthetic data, formulate or prove mathematical claims, or propose our core hypotheses, methodology, or experimental design; these were carried out by the authors. We have reviewed all AI-assisted work: code was checked against expected behavior and the underlying data, and all AI-drafted text and result interpretations were verified against experimental outputs and revised where necessary. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses sensitive identity verification data from a real-world industrial system. All data are handled under the firm's privacy and security requirements. Due to their sensitive nature, the full dataset cannot be released; only a limited, appropriately processed subset may be made public.

REPRODUCIBILITY STATEMENT

To advance the reproducibility of this work, we plan to release all training and evaluation scripts, configuration files, and model weights used in our experiments upon acceptance. Due to privacy and confidentiality concerns associated with user data from the production identity verification system, we cannot release the complete dataset. Instead, we plan to release a carefully selected subset of de-identified data samples under appropriate usage requirements to facilitate independent evaluation and further research.

REFERENCES

Shuai Bai et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Nuria Alina Chandra, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, Jongwook Choi, Aerin Kim, and Oren Etzioni. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. *arXiv preprint arXiv:2503.02857*, 2025.

Nuria Alina Chandra, Hannah Lee, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Changyeon Lee, Jongwook Choi, Sejin Paik, Aerin Kim, and Oren Etzioni. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 10668–10678, June 2026.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1251–1258, 2017.

Davide Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. *arXiv preprint arXiv:2107.02612*, 2022.

Christian Gerlach and Nicolas Poirel. Who's got the global advantage? visual field differences in processing of global and local shape. *Cognition*, 195:104131, 2020. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2019.104131. URL https://www.sciencedirect.com/science/article/pii/S0010027719303051.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Xiao Guo, Xiufeng Song, Yue Zhang, Xiaohong Liu, and Xiaoming Liu. Rethinking vision-language model in face forensics: Multi-modal interpretable forged face detector. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–116. IEEE, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Zhengchao Huang et al. Ffaa: Multimodal large language model based explainable open-world face forgery analysis assistant. *arXiv preprint arXiv:2408.10072*, 2024.

Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics. *arXiv preprint arXiv:2404.07966*, 2024.

Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2889–2898, 2020.

Ning Jiang, Dingheng Zeng, Yanhong Liu, Haiyang Yi, Shijie Yu, Minghe Weng, Haifeng Shen, and Ying Li. Explainable deepfake detection with rl enhanced self-blended images. *arXiv preprint arXiv:2601.15624*, 2026.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. URL https://arxiv.org/abs/1812.04948.

Tianxiao Li, Zhenglin Huang, Haiquan Wen, Yiwei He, Xinze Li, Bingyu Zhu, Wuhui Duan, Congang Chen, Zeyu Fu, Yi Dong, Baoyuan Wu, Xiangtai Li, and Guangliang Cheng. Omni-fake: Benchmarking unified multimodal social media deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30299–30311, June 2026.

Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7, 2018.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 772–781. IEEE, 2021a.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 772–781, 2021b.

Jiawei Liu et al. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*, 2024.

Scott McCloskey and Michael Albright. Detecting gan-generated imagery using color cues, 2018. URL https://arxiv.org/abs/1812.08247.

David Navon. Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, 9(3):353–383, 1977. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(77)90012-3. URL https://www.sciencedirect.com/science/article/pii/0010028577900123.

Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kulbe, Ehsan Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17–26, 2024.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models, 2024. URL https://arxiv.org/abs/2302.10174.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 86–103, 2020.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error, 2024. URL https://arxiv.org/abs/2401.17879.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Yichen Shi, Yuhao Gao, Yingxin Lai, Hongyang Wang, Jun Feng, Lei He, Jun Wan, Changsheng Chen, Zitong Yu, and Xiaochun Cao. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *Visual Intelligence*, 3(1):9, 2025.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18720–18729, 2022.

Sumsub. Identity fraud report 2024. 2024. Available at https://sumsub.com/fraud-report-2024.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114, 2019.

Yiming Tang, Abhijeet Sinha, and Dianbo Liu. How does my model fail? automatic identification and interpretation of physical plausibility failure modes with matryoshka transcoders, 2025. URL https://arxiv.org/abs/2511.10094.

Yiming Tang, Arash Lagzian, Srinivas Anumasa, Qiran Zou, Yingtao Zhu, Ye Zhang, Trang Nguyen, Yih-Chung Tham, Ehsan Adeli, Ching-Yu Cheng, Yilun Du, and Dianbo Liu. Human-like content analysis for generative ai with language-grounded sparse encoders, 2026. URL https://arxiv.org/abs/2508.18236.

Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(80)90005-5. URL https://www.sciencedirect.com/science/article/pii/0010028580900055.

Gaojian Wang, Feng Lin, Tong Wu, Zhenguang Liu, Zhongjie Ba, and Kui Ren. Fsfm: A generalizable face security foundation model via self-supervised facial representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24364–24376, 2025.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now, 2020. URL https://arxiv.org/abs/1912.11035.

Youqi Wang et al. Forgeryvcr: Visual-centric reasoning via efficient forensic tools in mllms for image forgery detection and localization. *arXiv preprint arXiv:2602.14098*, 2026.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection, 2023. URL https://arxiv.org/abs/2303.09295.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-ring watermarks: Fingerprints for diffusion images that are invisible and robust, 2023. URL https://arxiv.org/abs/2305.20030.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms, 2023. URL https://arxiv.org/abs/2312.14135.

Junhao Xu et al. Identity-aware vision-language model for explainable face forgery detection. *arXiv preprint arXiv:2504.09439*, 2025.

Wenbo Xu et al. Mare: Multimodal alignment and reinforcement for explainable deepfake detection via vision-language models. *arXiv preprint arXiv:2601.20433*, 2026.

Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. *arXiv preprint arXiv:1811.00661*, 2019.

Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints, 2019. URL https://arxiv.org/abs/1811.08180.

Peipeng Yu, Jianwei Fei, Hui Gao, Xuan Feng, Zhihua Xia, and Chip Hong Chang. Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72925–72943. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/yu25d.html.

Zitong Yu, Ajian Liu, Chenxu Zhao, Kevin H. M. Cheng, Xu Cheng, and Guoying Zhao. Flexible-modal face anti-spoofing: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 6346–6351, June 2023.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *Proceedings of the European Conference on Computer Vision (ECCV).*, 2024.

Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2185–2194, 2021.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Binjia Zhou et al. Evolvereason: Self-evolving reasoning paradigm for explainable deepfake facial image identification. *arXiv preprint arXiv:2603.07515*, 2026.

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, 2023. URL https://arxiv.org/abs/2306.08571.

Muzhi Zhu, Hao Zhong, Canyu Zhao, Zongze Du, Mingyu Liu, Zheng Huang, Anzhou Li, Hao Chen, Cheng Zou, Jingdong Chen, Ming Yang, and Chunhua Shen. Active-o3: Empowering mllms with active perception via pure reinforcement learning, 2026. URL https://arxiv.org/abs/2505.21457.

A DEPLOYMENT

ForensicZoom has been deployed in a production identity verification system serving millions of daily requests. It operates within a two-tier architecture: lightweight online models with approximately 0.5B parameters process all traffic at sub-100ms latency and classify requests as clearly legitimate, clearly fraudulent, or uncertain, while ForensicZoom processes only the uncertain gray-zone cases asynchronously. This design concentrates the computational cost of the MLLM on the ambiguous cases where additional forensic reasoning is most valuable.

Among gray-zone samples processed by ForensicZoom, approximately 23% trigger the zoom mechanism, while the remaining 77% are resolved in a single inference round. This rate is higher than the 17% observed on the full test set because the first-tier models have already filtered out most straightforward cases. Average end-to-end latency is 1.2s for single-round inference and 2.8s for zoom-triggered cases. Before deployment, manual review required an average of 5.3 minutes per case; with ForensicZoom's attribution reports, this is reduced to 1.8 minutes, corresponding to a 2.9$\times$ improvement in review efficiency. In production, newly emerging attack patterns are incorporated through incremental updates of Stages 3–4 on a two-week cycle without retraining Stages 1–2 from scratch.

B LIMITATIONS AND FUTURE WORK

A broader challenge facing the community is the absence of large-scale, multimodal identity verification benchmarks that reflect real-world deployment conditions. Existing public benchmarks provide isolated visual inputs and omit the behavioral signals—interaction timing, repeated verification attempts, device metadata—that are critical for reliable detection at operationally meaningful false-positive rates. Our use of proprietary production data is a direct response to this gap; we plan to release the training framework and a limited evaluation subset where permitted, and encourage the community to develop richer multimodal benchmarks.

The zoom mechanism introduces additional inference latency, mitigated in practice by our two-tier architecture but relevant for strictly real-time settings. Future work will investigate multi-scale and multi-round inspection, more efficient adaptation to emerging attack patterns, temporal reasoning for video-based verification, and evaluation on public benchmarks as they mature.