TRACER: DUAL EVIDENCE MODELING FOR GENERALIZABLE IMAGE FORGERY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in image generation and editing have led to increasingly diverse and realistic image forgeries, making reliable detection more challenging. However, existing forensic methods are typically designed for specific forgery forms and rely on type-specific artifacts, such as boundary and noise inconsistencies in locally manipulated images or structural and frequency irregularities in AI-generated images, making them difficult to extend to other forgery scenarios. To address this limitation, we propose **TRACER**, a generalizable image forgery detection framework for diverse forgery forms. Our key insight is that forgery processes introduce anomalous deviations from real-image distributions, which can be primarily characterized from two views, fine-grained and logical. Accordingly, TRACER adopts a dual-stream architecture that models each evidence view with a dedicated stream. To promote their collaboration, Bidirectional Evidence Sharing facilitates information exchange between streams, while Hybrid Layer Interaction controls where such exchange occurs across feature depths. Progressive Evidence Training then gradually activates cross-stream interaction. Across benchmarks covering several forgery scenarios, TRACER achieves state-of-the-art performance, improving average evidence localization mIoU by 8.3% and judgment accuracy by 22.5%. It also reaches 92.6% zero-shot accuracy on synthetic images and videos from recent generators. These results show the potential of dual evidence modeling for accurate and generalizable image forgery detection.

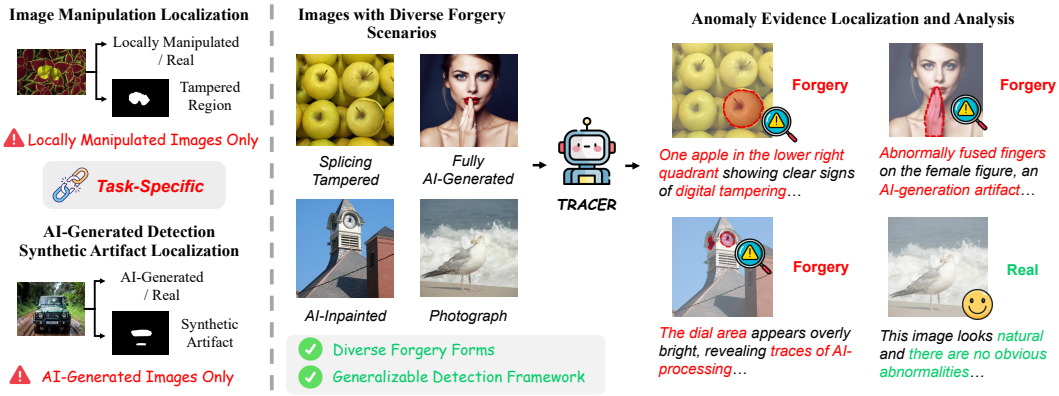
1 INTRODUCTION

Recent advances in image generation models (OpenAI, 2026b) and editing techniques (Wang et al., 2025) have broadened the forms of image forgery. Beyond conventional splicing (Kwon et al., 2022), modern forgeries may involve generative editing (Wang et al., 2025), full synthesis (OpenAI, 2026b), or combinations of these techniques. These advances have made forged content easier to produce but harder to verify, increasing risks of misinformation, fraud, and erosion of trust in digital media. Therefore, a unified detection framework must handle image forgeries produced through diverse forgery processes, while providing concrete and intuitive reasons for authenticity assessment.

Existing methods typically target specific forgery forms and formulate them as separate detection tasks. Image manipulation localization methods (Guillaro et al., 2023; Kwon et al., 2025) focus on anomalies within tampered regions of locally manipulated images, such as manipulation traces reflected in pixel-level inconsistencies. AI-generated image detectors (Wang et al., 2020; Karageorgiou et al., 2025) capture synthetic-specific artifacts to distinguish generated images from real ones, often relying on image-level classification. To improve interpretability, recent work (Kang et al., 2025) further makes such cues explicit by localizing visual artifacts or logical inconsistencies in fully AI-generated images. Recent unified detection methods (Huang et al., 2025; Du et al., 2026) attempt to handle multiple forgeries, but remain dependent on predefined categories or domain-specific anomalies, limiting their generalization beyond known forgery processes. As summarized in Fig. 1(a), these methods capture anomalies characteristic of particular forgery types, limiting their generalization to diverse and previously unseen generation and manipulation processes.

To achieve more generalizable detection, we seek a common perspective by focusing on the intrinsic properties shared across forgery processes. Without relying on predefined forgery types, we formulate detection as a unified question: *Does this image contain anomaly evidence suggesting that it*

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.



(a) Existing Methods (b) Our Image Forgery Evidence Localization and Analysis Method

Figure 1: Comparison of image forgery detection frameworks. (a) Existing methods are typically designed for predefined forgery types. (b) TRACER provides a unified framework that jointly localizes anomaly evidence and performs authenticity assessment in images from diverse forgery scenarios.

may be forged? Despite the diversity of forgery processes, forged images consistently introduce anomalous deviations from the distribution of authentic images, which serve as shared evidence for detection. Based on this observation, we formulate **Image Forgery Evidence Localization and Analysis**, a new task for identifying and analyzing forgery-related anomaly evidence to enable more generalizable detection across diverse forgery forms.

However, anomaly evidence can appear at different levels and spatial scales, making it difficult to model within a single framework. We group such evidence into two main views, as illustrated in Fig. 2: (i) *fine-grained evidence*, including unnatural boundaries, texture discontinuities, and noise mismatches, (ii) *logical evidence*, such as implausible relations, distorted structures and physical inconsistencies. The former is often subtle and requires dense spatial modeling, whereas the latter requires broader contextual reasoning about objects and scene relations. These evidence views capture different aspects of forgery anomalies and are not completely independent. Their inherent relationship can provide additional context across views and improve evidence representation. Therefore, detection requires not only dedicated modeling of each evidence view, but also mechanisms to leverage the relationship between them.

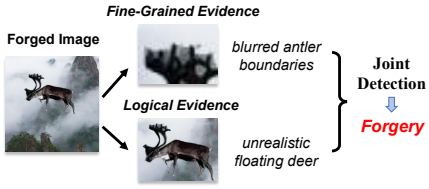


Figure 2: Joint detection of forged images from two anomaly evidence views.

To this end, we propose ***TRACER***, an end-to-end dual-stream framework with a fine-grained evidence stream and a logical evidence stream, each capturing anomalies from its corresponding view. To support more comprehensive evidence modeling, cross-stream modules exchange information at selected feature levels, while training gradually strengthens the interaction between evidence streams. Finally, the framework produces an evidence mask together with an authenticity judgment and analysis. Extensive experiments across diverse forgery forms show that TRACER achieves state-of-the-art performance and strong zero-shot generalization.

- We formulate *Image Forgery Evidence Localization and Analysis*, an evidence-centered task that supports image detection across diverse forms of forgery.
- We propose *TRACER*, a dual-stream framework that captures fine-grained and logical evidence. *Bidirectional Evidence Sharing* supports information exchange between the two streams, while *Hybrid Layer Interaction* applies this exchange at selected feature levels.
- We introduce *Progressive Evidence Training*, which gradually activates cross-stream interaction and then strengthens the model's localization and analysis capabilities.
- Extensive experiments across diverse forgery scenarios show that TRACER achieves state-of-the-art performance. It improves average localization mIoU by 8.3% and authenticity judgment accuracy by 22.5%, achieving 85.9% accuracy on images generated by GPT Image 2, showing strong generalization to unseen generators.

2 Related Work

Image Manipulation Detection and Localization. Image manipulation detection and localization aim to determine whether an image has been locally altered and to locate manipulated regions. Early methods target specific traces, such as copy-move duplication (He et al., 2023) or JPEG inconsistencies (Kwon et al., 2022), while later approaches combine noise, boundary, and content cues for broader manipulation localization (Chen et al., 2021; Guillaro et al., 2023). Recent methods extend beyond binary masks, with SAFIRE (Kwon et al., 2025) groups pixels by source for richer forgery analysis, while FakeShield (Xu et al., 2025) and ForgeryGPT (Zhang et al., 2024) further provide visual explanations. However, these methods remain tailored to local manipulation and centered on capturing manipulation-specific traces, limiting transfer to fully synthetic images.

AI-Generated Image Detection and Localization. AI-generated image detection is commonly formulated as a whole-image classification task that determines whether an image is real or synthetic. Early methods exploit generator artifacts (Wang et al., 2020) or diffusion reconstruction errors (Wang et al., 2023), while later work targets more general visual and statistical anomalies for unseen generators (Ojha et al., 2023; Yan et al., 2025; Karageorgiou et al., 2025). FakeVLM (Wen et al., 2025) and AIGI-Holmes (Zhou et al., 2025) provide textual explanations, helping users assess the supporting evidence. LEGION (Kang et al., 2025) further proposes to segment anomalous artifact regions caused by image synthesis. This approach makes the evidence for synthetic image detection explicit and visually interpretable. However, because these methods are designed for fully synthetic images, they are not well suited to locally manipulated or generatively edited images.

Unified Detection Across Diverse Forgery Forms. Only a few studies have explored unified detection across diverse forgery forms. SIDA (Huang et al., 2025) explores multi-class forgery recognition, but remains constrained by predefined categories and localizes only tampered regions. SICA (Du et al., 2026) unifies judgment across forensic domains, but only provides image-level real/fake scores without explicit localization or explanation. UniShield (Huang et al., 2026) uses domain routing and specialized experts, but its unification remains dependent on predefined domains and specialized models, with inconsistent capabilities. More broadly, these approaches still tend to rely on anomalies specific to particular forgery types, rather than broadly shared forgery anomaly. Overall, there remains a lack of a generalizable framework that jointly performs authenticity judgment, localization, and analysis across diverse and unknown forgery processes.

3 Method

In this section, we present TRACER, a dual-stream framework for image forgery evidence localization and analysis. We first define the task formulation. Then We describe the overall architecture, followed by Bidirectional Evidence Sharing (BES) and Hybrid Layer Interaction, which govern how and where information is exchanged between the streams. Finally, we introduce Progressive Evidence Training to gradually activate cross-stream interaction during training.

3.1 Task Definition

Given an image $I \in \mathbb{R}^{H\times W\times 3}$ that may have undergone unknown manipulations, the task is to localize anomaly evidence while producing a natural-language response comprising an authenticity judgment and anomaly analysis. The task is formally defined as follows:

$$(\hat{\mathbf{M}}, \hat{\mathbf{r}}) = F_\theta(I), \tag{1}$$

where $\hat{\mathbf{M}}$ is the predicted mask, $\hat{\mathbf{r}}$ is the generated judgment and analysis, F_θ is the model.

Together, $\hat{\mathbf{M}}$ and $\hat{r}$ provide a comprehensive forgery assessment from spatial and semantic perspectives. Specifically, $\hat{\mathbf{M}}$ highlights visually perceptible anomalous regions, including manipulated regions (e.g., splicing) and regions exhibiting physical inconsistencies or implausible relations (e.g., full AI-generation). For authentic images, $\hat{\mathbf{M}}$ approximates an all-zero mask. Complementarily, $\hat{r}$ provides a holistic authenticity judgment and describes broader or subtler inconsistencies that are difficult to represent spatially. In this way, our formulation supports generalizable detection across diverse forms of forgery, rather than being tailored to a single forgery form.

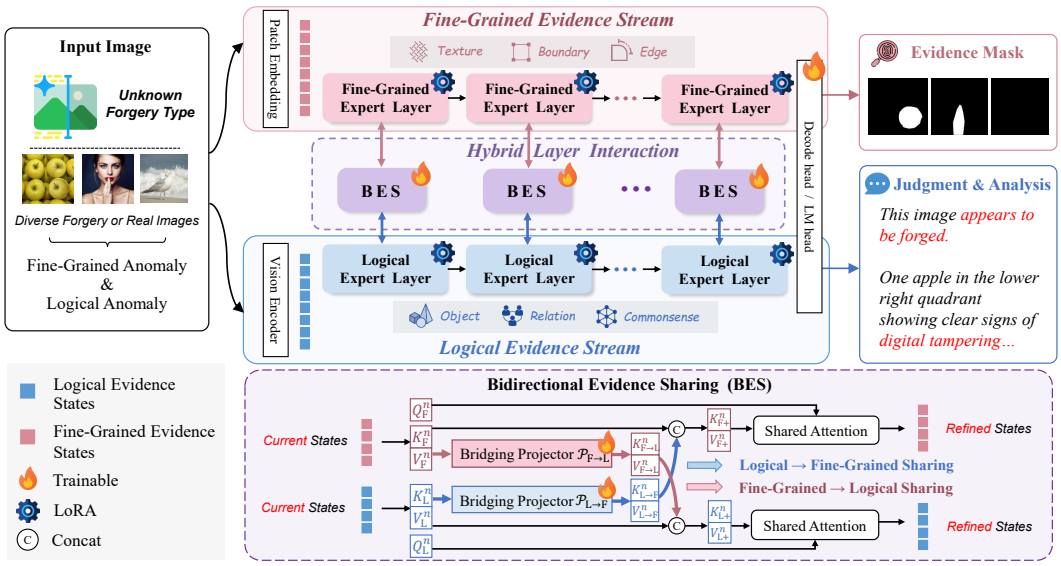


Figure 3: Overview of the proposed TRACER framework. TRACER contains a fine-grained evidence stream and a logical evidence stream. BES enables information exchange by allowing each stream to incorporate projected key-value states from the other, while Hybrid Layer Interaction determines the layer pairs at which this exchange occurs.

3.2 Architecture Overview

As shown in Fig. 3, TRACER consists of a fine-grained evidence stream and a logical evidence stream, each comprising a sequence of evidence expert layers. The fine-grained evidence stream employs SAM-based fine-grained experts (Kirillov et al., 2023), which provide dense spatial representations and are intended to model local forensic cues, such as boundary artifacts and texture inconsistencies. The logical evidence stream uses VLM-based logical experts (Bai et al., 2025) to obtain contextual representations of objects and their relations, capturing anomalies such as distorted structures and implausible interactions. To support more comprehensive evidence modeling, TRACER uses Bidirectional Evidence Sharing to exchange information between the two streams. Hybrid Layer Interaction determines the feature depths at which information is exchanged. After the final interaction layer, the model outputs the evidence mask and the language response containing the authenticity judgment and analysis.

3.3 Bidirectional Evidence Sharing

To enable the two streams to collaboratively model forgery evidence while preserving each stream's distinct objective, we introduce Bidirectional Evidence Sharing (BES) to exchange information between layers of the two streams. Specifically, each stream retains its own queries while attending to both its native key-value states and the projected key-value states from the other stream, enriching its attention context. As the two streams capture different evidence views, each gains richer context to model its own evidence. This provides the fine-grained stream with broader semantic information, while the logical stream gains access to localized visual details.

At an interaction layer n, let H_{F}^n and H_{L}^n denote the hidden states entering this layer from the fine-grained stream and the logical stream, respectively. Their query, key, and value states are denoted as $(Q_{\mathrm{F}}^n, K_{\mathrm{F}}^n, V_{\mathrm{F}}^n)$ and $(Q_{\mathrm{L}}^n, K_{\mathrm{L}}^n, V_{\mathrm{L}}^n)$. BES employs two learnable bridging projectors $\mathcal{P}_{\mathrm{L}\to\mathrm{F}}$ and $\mathcal{P}_{\mathrm{F}\to\mathrm{L}}$ to map the key-value states of each stream into the hidden space of the other:

$$(K_{\mathrm{L}\to\mathrm{F}}^n, V_{\mathrm{L}\to\mathrm{F}}^n) = \mathcal{P}_{\mathrm{L}\to\mathrm{F}}(K_{\mathrm{L}}^n, V_{\mathrm{L}}^n), \quad (K_{\mathrm{F}\to\mathrm{L}}^n, V_{\mathrm{F}\to\mathrm{L}}^n) = \mathcal{P}_{\mathrm{F}\to\mathrm{L}}(K_{\mathrm{F}}^n, V_{\mathrm{F}}^n). \tag{2}$$

The bridged states are then concatenated with the native key-value states of the target stream:

$$K^n_{\mathrm{F},+} = [K^n_{\mathrm{F}}; K^n_{\mathrm{L}\to\mathrm{F}}], \quad V^n_{\mathrm{F},+} = [V^n_{\mathrm{F}}; V^n_{\mathrm{L}\to\mathrm{F}}], \tag{3}$$

$$K^n_{\mathrm{L},+} = [K^n_{\mathrm{L}}; K^n_{\mathrm{F}\to\mathrm{L}}], \quad V^n_{\mathrm{L},+} = [V^n_{\mathrm{L}}; V^n_{\mathrm{F}\to\mathrm{L}}]. \tag{4}$$

Each stream feeds its native query and the merged key-value states into Shared Attention to obtain the updated hidden state:

$$\tilde{H}^n_{\mathrm{F}} = \mathrm{SharedAttn}(Q^n_{\mathrm{F}}, K^n_{\mathrm{F},+}, V^n_{\mathrm{F},+}), \quad \tilde{H}^n_{\mathrm{L}} = \mathrm{SharedAttn}(Q^n_{\mathrm{L}}, K^n_{\mathrm{L},+}, V^n_{\mathrm{L},+}). \tag{5}$$

The updated states $\tilde{H}^n_{\mathrm{F}}$ and $\tilde{H}^n_{\mathrm{L}}$ are passed to the next layers of their corresponding streams. With this design, BES allows each stream to keep its own evidence extraction path while incorporating key-value states from the other stream to enhance its own representation. As a result, each stream gains richer context for modeling its respective evidence.

3.4 Hybrid Layer Interaction

Applying BES densely across many layers allows the two streams to interact frequently, but increases computational cost and may cause excessive coupling between their representations. In contrast, overly sparse interaction may limit information exchange, while interaction at a single depth may miss relevant features represented at other levels. Therefore, we introduce Hybrid Layer Interaction to balance interaction frequency and depth coverage. It applies BES to selected layer pairs distributed across depths while keeping the remaining layers independent. Since the two streams may have different numbers of layers, the selected pairs are aligned according to their relative depths.

Let N_{F} and N_{L} denote the numbers of layers in the fine-grained and logical streams, respectively, and let i^{F}_k and i^{L}_k denote the selected layer indices for the k-th interaction in the two streams. Using zero-based indexing, we select every third layer of the logical expert for interaction, starting from index 2, and select the corresponding fine-grained expert layers by aligning their relative depths:

$$i^{\mathrm{L}}_k = 3k + 2, \quad i^{\mathrm{F}}_k = \left\lfloor \frac{N_{\mathrm{F}}}{N_{\mathrm{L}}} i^{\mathrm{L}}_k \right\rfloor, \quad k = 0, \ldots, \left\lfloor \frac{N_{\mathrm{L}}}{3} \right\rfloor - 1. \tag{6}$$

At each selected layer pair, BES jointly updates the states of the two streams:

$$(\tilde{H}^{i^{\mathrm{F}}_k}_{\mathrm{F}}, \tilde{H}^{i^{\mathrm{L}}_k}_{\mathrm{L}}) = \mathrm{BES}(H^{i^{\mathrm{F}}_k}_{\mathrm{F}}, H^{i^{\mathrm{L}}_k}_{\mathrm{L}}). \tag{7}$$

All nonselected layers continue independently within their respective streams. This schedule enables information exchange at different network depths while allowing most layers in each stream to retain their own modeling capabilities, achieving a balance between interaction and efficiency.

3.5 Progressive Evidence Training

The two evidence streams are initialized from backbones trained for different objectives and may thus exhibit different feature distributions at the beginning of training. Enabling interaction at full strength from the outset may lead to unstable optimization before the streams have adapted to the target task, potentially resulting in suboptimal training outcomes. Therefore, we propose Progressive Evidence Training, which gradually increases the strength of interaction between the streams, and then sequentially fine-tunes the model to strengthen its capabilities in analysis and localization.

Stage 1: Cross-Evidence Gated Alignment. At each BES layer, a scheduled bias is applied to the attention logits of bridged keys, gradually gradually increases cross-stream contributions while leaving native-key logits unchanged. With warm-up length T_w, the gate value at training step t is:

$$\gamma(t) = \frac{1}{2}\left[1 - \cos\left(\pi \cdot \min\left(1, \frac{t}{T_w}\right)\right)\right], \tag{8}$$

Both streams are updated in parallel. For either direction, let x denote the stream being updated and y the other stream, where $(x, y) \in \{(\mathrm{F}, \mathrm{L}), (\mathrm{L}, \mathrm{F})\}$. The gate introduces a bias of $\log(\epsilon + \gamma(t))$ to

bridged-key logits while leaving native-key logits unchanged. The gated BES update for x is:

$$\tilde{H}_x^n = \mathrm{Softmax}\left(\frac{Q_x^n (K_{x,+}^n)^\top}{\sqrt{d}} + [\mathbf{0}, \log\left(\epsilon + \gamma(t)\right)\mathbf{1}]\right) V_{x,+}^n. \tag{9}$$

where ϵ is a small positive constant for numerical stability. The first block leaves native-key logits unchanged, while the second applies the scheduled bias to bridged-key logits.

The same bias is applied to the attention logits of every query. Early in training, the negative bias suppresses attention to bridged keys. As $\gamma(t)$ approaches 1, the bias becomes approximately zero, gradually increasing cross-stream interaction while leaving native-key logits unchanged.

Stage 2: Capability Refinement. At this stage, we perform supervised fine-tuning, first for authenticity judgment and analysis, then for evidence localization, to refine both capabilities.

Training Objective. We use the following objective throughout training:

$$\mathcal{L} = \lambda_{\text{txt}}\mathcal{L}_{\text{txt}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}, \tag{10}$$

where $\mathcal{L}_{\text{seg}}$ denotes the segmentation loss combining Dice and binary cross-entropy for evidence localization, and $\mathcal{L}_{\text{txt}}$ the language-generation loss for authenticity analysis.

4 Experiments

Implementation Details. We build TRACER upon SAM (Kirillov et al., 2023) as the fine-grained evidence stream and Qwen3-VL-4B (Bai et al., 2025) as the logical evidence stream. Training runs on 8 NVIDIA RTX 4090 GPUs. Detailed experimental settings are provided in the Appendix.

Training Datasets. We train TRACER on 180K samples from CASIA v2 (Dong et al., 2013), FantasticReality (Kniaz et al., 2019), IMD2020 (Novozamsky et al., 2020), tampered COCO (Lin et al., 2014), tampered RAISE (Dang-Nguyen et al., 2015), SynthScars train set (Kang et al., 2025), and FakeClue (Wen et al., 2025). These datasets cover diverse forgery scenarios with localization supervision. We also use Qwen3-VL-235B (Bai et al., 2025) to generate detailed explanations for each image. Detailed training configuration and data construction are provided in the Appendix.

Benchmarks. We evaluate TRACER under three representative forgery settings: (i) traditional local manipulation on CASIAv1+ (Chen et al., 2021), Coverage (Wen et al., 2016), NIST16 (Guan et al., 2019), and Columbia (Hsu & Chang, 2006), (ii) generative editing, including deepfakes and AI inpainting, on OpenForensics (Le et al., 2021) and CocoGlide (Guillaro et al., 2023), (iii) full AI-generation on SynthScars test set (Kang et al., 2025), LOKI (Ye et al., 2025), and RichHF-18K (Liang et al., 2024), all with human-annotated anomaly regions. We also evaluate false-positive attribution on real images from COCO 2017 (Lin et al., 2014) and SITD (Chen et al., 2018). ***These benchmarks are well-established, high-quality, and challenging, and have been widely used in prior work*** (Guillaro et al., 2023; Kang et al., 2025). To assess generalization, we report results on OpenFake (Livernoche et al., 2025), comprising outputs from recent image generators such as GPT Image 2 (OpenAI, 2026b), Nano Banana Pro (Google DeepMind, 2025), and Seedream 5.0 (ByteDance Seed Team, 2026), as well as frames from video generators such as Sora 2 (OpenAI, 2025) and Wan 2.5 (Alibaba Cloud, 2025).

Evaluation Metrics. Following prior work (Guillaro et al., 2023), we evaluate forgery evidence localization using mIoU measuring mask overlap, while F1 score reflects the precision–recall balance of localized regions. We use text accuracy (Acc) to evaluate whether the analysis correctly judges the image's authenticity, and report an Explanation Quality Score (EQS; 0–10), assigned by Claude Sonnet 5 (Anthropic, 2026) as an MLLM judge (Chen et al., 2024b), to assess the quality of anomaly analysis. Details are provided in the appendix.

4.1 Main Results

Forgery Evidence Localization. As shown in Table 1, we evaluate TRACER's localization performance across three representative forgery scenarios: traditional local manipulation, generative

editing, and full AI-generation. Existing methods emphasize specific evidence, such as frequency artifacts or semantic anomalies, making it difficult to generalize their localization capabilities across diverse forms of forgery. TRACER addresses this limitation by integrating fine-grained and logical evidence to capture distributional discrepancies between forged and real images, enabling anomaly localization across diverse forgery scenarios. It improves average mIoU and F1 by 8.3% and 9.7 % over TruFor (Guillaro et al., 2023), which exploits noise inconsistencies, and also outperforms SIDA (Huang et al., 2025), which models semantic anomalies. Notably, CocoGlide (Guillaro et al., 2023) is a challenging benchmark for generative editing, where manipulated regions blend naturally into their surroundings, obscuring manipulation boundaries. TRACER's substantial gains suggest that dual-evidence modeling helps localize these difficult-to-distinguish, well-blended forgeries.

Table 1: Comparison of forgery evidence localization across three representative forgery scenarios using mIoU and F1 as evaluation metrics. **Bold** indicates the best performance, and underlined denotes the second-best. **For fair comparison, baselines are all trained on the same training data as our method, ensuring a consistent training data setting across all methods.**

Method	Trad. Local Manipulation						Generative Edit				Fully AI-generated						AVE	
	CASIAv1+		Coverage		NIST16		OpenFor.		CocoGlide		SynthScars		LOKI		RichHF			
	mIoU	F1	mIoU	F1	mIoU	F1	mIoU	F1	mIoU	F1	mIoU	F1	mIoU	F1	mIoU	F1	mIoU	F1
MVSS-Net (Chen et al., 2021)	.695	.503	.583	.404	.526	.176	.788	.791	.544	.468	.547	.324	.440	.284	.519	.231	.580	.398
DetectiveSAM (Lek et al., 2026)	.608	.372	.511	.266	.476	.124	.552	.373	.521	.485	.494	.149	.435	.262	.482	.059	.510	.261
TruFor (Guillaro et al., 2023)	.753	.685	.626	.513	.509	.290	**.829**	.827	.512	.475	.538	.314	.442	**.304**	.520	**.232**	.591	.455
CAT-Net v2(Kwon et al., 2022)	.753	.718	.606	.470	.523	.287	.746	.641	.610	.502	.518	.210	**.447**	.284	.506	.066	.589	.397
SAFIRE(Kwon et al., 2025)	.471	.315	.441	.401	.531	.440	.431	.243	.436	.478	.422	.217	.407	.204	.419	.101	.445	.300
Mesorch(Zhu et al., 2025)	.853	.821	.614	.479	.546	.208	.774	.785	.518	.476	.503	.257	.397	.234	.475	.056	.585	.415
LISA(7B)(Lai et al., 2024)	.721	.534	.507	.307	.559	.268	.705	.622	.530	.500	.524	.273	.436	.269	.523	.142	.563	.364
FakeShield(22B)(Xu et al., 2025)	.724	.520	.476	.233	.548	.292	.633	.485	.543	.506	.377	.189	.444	.268	.372	.164	.515	.332
LEGION(8B)(Kang et al., 2025)	.653	.439	.464	.217	.527	.257	.629	.443	.534	.481	.525	.226	.440	.284	.503	.159	.534	.313
SIDA(13B)(Huang et al., 2025)	.693	.511	.492	.268	.537	.213	.679	.591	.554	.503	.534	.278	.443	.285	.515	.163	.556	.352
TRACER(4B)	**.919**	**.893**	**.830**	**.760**	**.650**	**.477**	.814	**.836**	**.655**	**.636**	**.560**	**.329**	.435	.287	**.532**	.196	**.674**	**.552**

Authenticity Judgment and Analysis. Table 2 compares TRACER with closed-source multimodal models such as GPT-5.5 (OpenAI, 2026a) and open-source models specialized in forgery detection, such as SIDA (Huang et al., 2025), on image authenticity judgment and anomaly analysis. Many baselines exhibit an imbalanced tendency to favor real-image predictions over forgery detection, resulting in high acceptance of real images but poor recognition of forgeries. In contrast, TRACER's dual-evidence modeling enables precise capture of anomalies in forged images, supporting accurate detection across diverse forgery types, fewer false positives on real images, and detailed explanations of its authenticity judgments. It achieves highly competitive Acc on forged images, outperforming most existing models. Moreover, TRACER achieves the highest balanced accuracy, surpassing Gemini 3.1 Pro (Google DeepMind, 2026) by 22.5%, and obtains the excellent EQS, indicating the strongest analysis capability in explaining forgery-related visual inconsistencies.

Table 2: Comparison of authenticity judgment and analysis across three forgery scenarios and real images using Acc and EQS. All results are averaged over three independent evaluation runs. **-** indicates that a model does not support the corresponding capability.

Method	Trad. Local Manipulation				Generative Edit				Fully AI-generated				Real			
	CASIAv1+		Columbia		OpenFor.		CocoGlide		SynthScars		LOKI		COCO		SITD	
	Acc	EQS	Acc	EQS	Acc	EQS	Acc	EQS	Acc	EQS	Acc	EQS	Acc	EQS	Acc	EQS
GPT-5.5 (OpenAI, 2026a)	.182	8.13	.789	8.82	.055	5.46	.006	5.39	.865	8.39	.524	6.38	.843	7.46	.911	8.03
Gemini 3.1 Pro (Google DeepMind, 2026)	.610	8.78	.902	8.17	.116	5.92	.012	4.76	.879	7.07	.576	7.63	.872	8.27	.900	8.12
Qwen3.7 Max (Qwen Team, 2026)	.187	4.67	.839	8.01	.024	6.79	.005	5.96	.815	5.51	.654	**8.96**	.713	8.34	.967	7.81
FakeShield(22B) (Xu et al., 2025)	.896	7.31	.448	7.09	.897	3.04	**.843**	8.23	-	-	-	-	.407	5.04	.944	6.77
FakeVLM(7B) (Wen et al., 2025)	-	-	-	-	.275	2.83	.027	3.54	.934	4.89	.661	4.65	**.884**	6.58	.144	6.90
SIDA(13B) (Huang et al., 2025)	.316	6.45	.655	5.46	.022	4.74	.553	2.99	.182	3.75	.057	2.04	.748	3.76	.833	5.04
TRACER(4B)	**.946**	**8.98**	**.995**	**9.14**	**.998**	**8.69**	.745	**8.61**	**.999**	**8.45**	**.842**	8.50	.874	**8.78**	**.989**	**8.39**

Zero-Shot Generalization to Recent Generators. We evaluate TRACER's out-of-domain (OOD) generalization on OpenFake (Livernoche et al., 2025), which covers images and videos from recent generators. Neither TRACER nor the baselines are trained on this dataset, and the highly realistic outputs of these generators further challenge their ability to capture diverse subtle anomalies. As shown in Table 3, TRACER achieves the highest average accuracy of authenticity judgment, reaching 85.9% on images generated by GPT Image 2 (OpenAI, 2026b) and outperforming FakeVLM (Wen et al., 2025) by 6.4%, which primarily relies on semantic cues for detection. By focusing on deviations from real-image distributions rather than generator-specific artifacts, TRACER captures diverse deviations from real image patterns rather than generator-specific artifacts, enabling robust generalization and reliable authenticity assessment on unseen generators.

Table 3: Comparison of zero-shot authenticity judgment using Acc to assess model generalization across images and videos generated by unseen image and video generators.

Method	Image			Video		AVE
	GPT Image 2	Nano Banana Pro	Seedream 5.0	Sora 2	Wan 2.5	
	Acc	Acc	Acc	Acc	Acc	Acc
CNNSpot (Wang et al., 2020)	.002	.000	.000	.000	.000	.000
UnivFD (Ojha et al., 2023)	.097	.117	.121	.108	.045	.098
AIDE (Yan et al., 2025)	.040	.184	.170	.072	.127	.119
Aligned (Rajan et al., 2025)	.032	.098	.876	.041	.008	.211
DRCT (Chen et al., 2024a)	.068	.681	.481	.266	.536	.406
SPAI (Karageorgiou et al., 2025)	.553	.360	.949	.386	.443	.538
CO-SPY (Cheng et al., 2025)	.249	.295	.433	.488	.528	.399
GAPL (Qin et al., 2026)	.597	.593	.860	.585	.761	.679
DDA (Chen et al., 2025)	.217	.878	.809	.505	.911	.664
SICA (Du et al., 2026)	.205	.334	.355	.418	.137	.290
SIDA(13B) (Huang et al., 2025)	.354	.414	.698	.438	.597	.500
FakeVLM(7B) (Wen et al., 2025)	.795	.890	.982	.804	.960	.886
TRACER(4B)	**.859**	**.933**	**.987**	**.869**	**.981**	**.926**

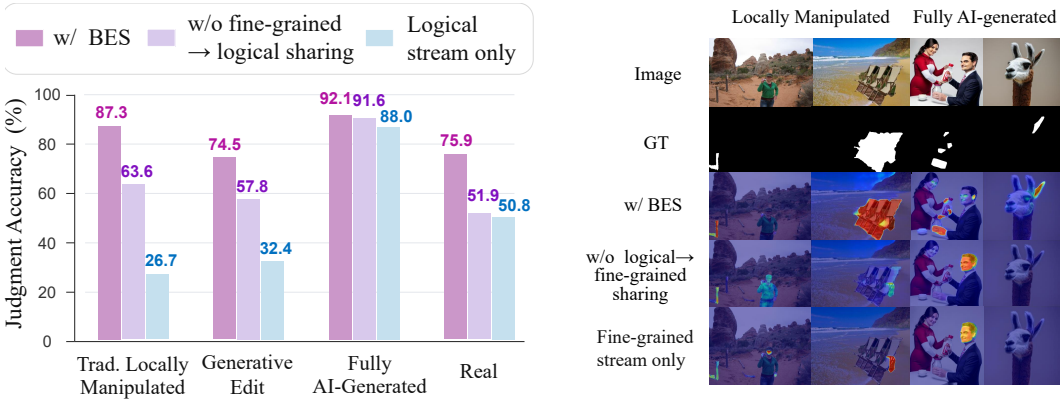
4.2 Ablation Studies

Bidirectional Evidence Sharing. As shown in Table 4, we evaluate the effectiveness of TRACER's dual-evidence modeling and bidirectional interaction. The single-stream variants underperform the dual-stream model, showing that neither evidence type alone is sufficient and that their interaction is important for diverse forgeries. Removing fine-grained→logical sharing causes a larger drop in authenticity judgment, particularly for local edits, whereas removing logical→fine-grained sharing mainly degrades localization, especially for fully synthetic images. With both directions enabled, BES achieves the highest average localization and judgment scores among the variants, supporting the benefit of bidirectional evidence interaction.

Qualitative results further clarify the respective roles of the two evidence views. In Fig. 4 (a), the logical stream variant shows the weakest separation between local edits and real images, while disabling fine-grained→logical sharing also noticeably reduces this separation. This indicates that localized forensic cues from the fine-grained stream help identify subtle anomalies and reduce false alarms. In Fig. 4 (b), both the fine-grained stream variant and removing logical→fine-grained sharing degrade the localization of semantic- and object-related anomalies, leading to missed evidence and spurious regions. This suggests that semantic context from the logical stream guides evidence localization. For example, in the first case, recognizing that two identical sticks violate the expected image content helps the model focus on the manipulated stick. Together, these results show the distinct roles of the two sharing directions.

Table 4: Ablation study of dual-evidence modeling and BES, validating the effectiveness of joint modeling and bidirectional interaction between fine-grained and logical evidence. Each single-stream variant supports only its corresponding function.

Setting	Trad. Local Manipulation				Gen. Edit		Fully AI-generated				AVE	
	CASIAv1+		Coverage		CocoGlide		SynthScars		LOKI			
	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc
Single Evidence Stream												
fine-grained stream only	.882	-	.757	-	.643	-	.481	-	.419	-	.636	-
logical stream only	-	.393	-	.140	-	.324	-	.971	-	.789	-	.523
Dual Evidence Stream												
w/o fine-grained→logical sharing	.914	.652	.829	.620	.646	.578	**.562**	.992	**.435**	.840	.677	.736
w/o logical→fine-grained sharing	.912	.940	.825	.760	.641	.674	.474	**.999**	.420	.836	.654	.842
w/ BES	**.919**	**.946**	**.830**	**.800**	**.655**	**.745**	.560	**.999**	**.435**	**.842**	**.680**	**.866**



(a) Effect of fine-grained evidence

(b) Effect of logical evidence

Figure 4: Effect of fine-grained evidence and logical evidence on authenticity judgment and localization for different forgery scenarios.

Hybrid Layer Interaction. Table 5 compares interaction schedules with different densities and depth distributions. Dense interaction increases latency without improving performance, whereas sparse interaction reduces both localization and judgment scores. Among schedules with the same number of interaction layers, distributing them across multiple depths yields higher scores than concentrating them at a single early, middle, or late stage. Under the evaluated settings, the hybrid schedule offers the most favorable performance-latency trade-off.

Table 5: Ablation study of Hybrid Layer Interaction with varying interaction densities and depth distributions. #Int. denotes the number of active cross-stream interaction layers, and latency denotes inference time in seconds.

Layer Sel.	#Int.	CASIAv1+		SynthScars		Latency
		mIoU	Acc	mIoU	Acc	
Interaction Density						
Sparse	4	.912	.912	.492	.789	10.04
Dense	32	.627	.681	.487	.992	15.47
Interaction Depth						
Early	12	.578	.650	.490	.994	10.73
Middle	12	.907	.864	.547	.989	10.73
Late	12	.505	.124	.461	.067	10.73
Hybrid	12	**.919**	**.946**	**.560**	**.999**	10.73

Table 6: Ablation study of the Progressive Evidence Training. Train denotes whether Stage 1 is performed, while Gate controls whether Stage 1 uses scheduled gating instead of direct training. Stage 2 corresponds to subsequent refinement.

Stage 1		Stage 2	CASIAv1+		SynthScars	
Train	Gate		mIoU	Acc	mIoU	Acc
✓	✗	✗	.903	.561	.534	.938
✓	✓	✗	.912	.879	.557	.982
✗	✗	✓	.854	.497	.491	.924
✓	✓	✓	**.919**	**.946**	**.560**	**.999**

Progressive Evidence Training. Comparing the first two rows of Table 6, the scheduled gate yields higher localization and analysis scores than directly enabling cross-stream interaction. Training with Stage 2 alone results in lower performance, suggesting that Stage 1 provides a useful initialization before capability refinement. Combining the scheduled Stage 1 with Stage 2 obtains the best results among the evaluated training configurations.

5 Conclusion

In this work, we propose Image Forgery Evidence Localization and Analysis as a new task. To address it, we introduce TRACER, a dual-stream framework that jointly captures fine-grained and logical anomaly evidence in forged images to enable generalizable detection across diverse forgery forms. Bidirectional Evidence Sharing and Hybrid Layer Interaction enable information exchange at selected feature levels, while Progressive Evidence Training gradually strengthens cross-stream interaction. Experiments across diverse forgery scenarios demonstrate state-of-the-art performance and strong zero-shot generalization. These results highlight TRACER's potential to generalize to emerging forgery techniques. We believe that evidence-centered modeling may enable more reliable and interpretable detection, safeguarding visual authenticity and trust in digital media.

AI USE STATEMENT

We use generative AI tools to suggest and refine wording in selected passages, improve English expression, identify potentially relevant references, and provide feedback on figure design and the presentation of experimental comparisons. We also used AI models to generate training explanations, which underwent the quality-control procedure described in Appendix A.1.1. We review and edit all AI-assisted material, independently verified the cited literature and technical claims, and take full responsibility for the final manuscript and artifacts. We do not use generative AI to formulate or prove mathematical claims.

REPRODUCIBILITY STATEMENT

The TRACER architecture, training objectives, datasets, and evaluation metrics are described in the main paper. Appendix A provides the data construction pipeline, detailed training configurations, computational resources, the exact layer-pairing scheme, and the prompts used for explanation quality evaluation. All experiments are repeated over three independent runs with different random seeds, each following the complete training and evaluation pipeline. We report the mean results across these runs. Code and trained models will be made publicly available.

REFERENCES

Alibaba Cloud. Commercial launch of Wan2.5-Preview. https://www.alibabacloud.com/en/events/wan2-5-preview-launch, November 2025. Accessed: 2026-09-07.

Anthropic. Claude Sonnet 5 system card. https://www.anthropic.com/claude-sonnet-5-system-card, June 2026. Accessed: 2026-09-26.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

ByteDance Seed Team. Deeper thinking, more accurate generation: Introducing Seedream 5.0 Lite. https://seed.bytedance.com/en/blog/deeper-thinking-more-accurate-generation-introducing-seedream-5-0-lite, February 2026. Accessed: 2026-09-07.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *International Conference on Machine Learning*, pp. 7621–7639, 2024a.

Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3291–3300, 2018.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *International Conference on Machine Learning*, pp. 6562–6595, 2024b.

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 *Advances in Neural Information Processing Systems*, 38: 106475–106500, 2025.

Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14185–14193, 2021.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. CO-SPY: Combining semantic and pixel features to detect synthetic images by AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13455–13465, 2025.

Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. RAISE: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM Multimedia Systems Conference*, pp. 219–224, 2015.

Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. MVSS-Net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2023.

Jing Dong, Wei Wang, and Tieniu Tan. CASIA image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pp. 422–426, 2013.

Bo Du, Xiaochen Ma, Xuekang Zhu, Zhe Yang, Chaoqun Niu, Chenfan Qu, Mingqi Fang, Zhenming Wang, Jingjing Liu, Jian Liu, and Ji-Zhe Zhou. Can we build a monolithic model for fake image detection? SICA: Semantic-induced constrained adaptation for unified-yet-discriminative artifact feature space reconstruction. In *International Conference on Machine Learning*, 2026.

Google DeepMind. Gemini 3 Pro Image model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Image-Model-Card.pdf, November 2025. Accessed: 2026-09-07.

Google DeepMind. Gemini 3.1 Pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-1-Pro-Model-Card.pdf, February 2026. Accessed: 2026-09-07.

Haiying Guan, Mark Kozak, Eric Robertson, Yooyoung Lee, Amy N. Yates, Andrew Delgado, Daniel Zhou, Timothee Kheyrkhah, Jeff Smith, and Jonathan Fiscus. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pp. 63–72, 2019.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. TruFor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20606–20615, 2023.

Yingjie He, Yuanman Li, Changsheng Chen, and Xia Li. Image copy-move forgery detection via deep cross-scale PatchMatch. In *IEEE International Conference on Multimedia and Expo*, pp. 2327–2332, 2023.

Yu-Feng Hsu and Shih-Fu Chang. Detecting image splicing using geometry invariants and camera characteristics consistency. In *IEEE International Conference on Multimedia and Expo*, pp. 549–552, 2006.

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 *International Conference on Learning Representations*, 2022.

Qing Huang, Zhipei Xu, Xuanyu Zhang, Xiangyu Yu, and Jian Zhang. UniShield: An adaptive multi-agent framework for unified forgery image detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 8121–8132, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28831–28841, 2025.

Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, and Conghui He. LEGION: Learning to ground and explain for synthetic image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18937–18947, 2025.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution AI-generated image detection by spectral learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18706–18717, 2025.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Vladimir V. Kniaz, Vladimir A. Knyaz, and Fabio Remondino. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. In *Advances in Neural Information Processing Systems*, 32:215–226, 2019.

Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning JPEG compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022.

Myung-Joon Kwon, Wonjun Lee, Seung-Hun Nam, Minji Son, and Changick Kim. SAFIRE: Segment any forged image region. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4437–4445, 2025.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Trung-Nghia Le, Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. OpenForensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10117–10127, 2021.

Gert Lek, Nicolas van Schaik, Chaoyi Zhu, Pin-Yu Chen, Robert Birke, and Lydia Y. Chen. Detective SAM: Adaptive AI-image forgery localization. In *International Conference on Learning Representations*, 2026.

Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy Dj Dvijotham, Katherine M. Collins, Yiwen Luo, Yang Li, Kai J. Kohlhoff, Deepak Ramachandran, and Vidhya Navalpakkam. Rich human feedback for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19401–19411, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Victor Livernoche, Akshatha Arodi, Andreea Musulan, Zachary Yang, Adam Salvail, Gaétan Marceau Caron, Jean-François Godbout, and Reihaneh Rabbany. OpenFake: An open dataset and platform toward real-world deepfake detection. *arXiv preprint arXiv:2509.09495*, 2025.

Adam Novozamsky, Babak Mahdian, and Stanislav Saic. IMD2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, pp. 71–80, 2020.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24480–24489, 2023.

OpenAI. Sora 2 system card. https://cdn.openai.com/pdf/50d5973c-c4ff-4c2d-986f-c72b5d0ff069/sora_2_system_card.pdf, September 2025. Accessed: 2026-09-07.

OpenAI. GPT-5.5 system card. https://deploymentsafety.openai.com/gpt-5-5/gpt-5-5.pdf, April 2026a. Accessed: 2026-09-07.

OpenAI. GPT Image 2: OpenAI API model documentation. https://developers.openai.com/api/docs/models/gpt-image-2, April 2026b. Accessed: 2026-09-07.

Ziheng Qin, Yuheng Ji, Renshuai Tao, Yuxuan Tian, Yuyang Liu, Yipu Wang, and Xiaolong Zheng. Scaling up AI-generated image detection with generator-aware prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43008–43017, 2026.

Qwen Team. Qwen3.7: The agent frontier. https://qwen.ai/blog?id=qwen3.7, May 2026. Accessed: 2026-09-07.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *International Conference on Learning Representations*, 2025.

Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. SeedEdit 3.0: Fast and high-quality generative image editing. *arXiv preprint arXiv:2506.05083*, 2025.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8695–8704, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023.

Bihan Wen, Ye Zhu, Ramanathan Subramanian, Tian-Tsong Ng, Xuanjing Shen, and Stefan Winkler. COVERAGE: A novel database for copy-move forgery detection. In *IEEE International Conference on Image Processing*, pp. 161–165, 2016.

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*, 38: 58972–59005, 2025.

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 *International Conference on Learning Representations*, 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 *International Conference on Learning Representations*, 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 *International Conference on Learning Representations*, 2025.

Fanrui Zhang, Jiawei Liu, Jiaying Zhu, Esther Sun, Dong Li, Qiang Zhang, and Zheng-Jun Zha. ForgeryGPT: A multimodal LLM for interpretable image forgery detection and localization. *arXiv preprint arXiv:2410.10238*, 2024.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18746–18758, 2025.

Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: Orchestrating multi-scale and hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11022–11030, 2025.

A Appendix

A.1 Training Details

A.1.1 Data Construction Pipeline

Each training sample consists of an input image, a localization mask for anomaly evidence regions, and an evidence-grounded textual analysis supporting the authenticity judgment. This common format provides unified supervision across the training set. The complete dataset will be made publicly available.

Step 1: Collect and standardize the data. We collect samples from CASIAv2 (Dong et al., 2013), FantasticReality (Kniaz et al., 2019), IMD2020 (Novozamsky et al., 2020), tampered COCO (Lin et al., 2014), tampered RAISE (Dang-Nguyen et al., 2015), and the SynthScars training split (Kang et al., 2025). Together, these datasets cover real images and diverse forgery scenarios, including traditional local manipulation, generative editing, and fully AI-generated images. Across these datasets, the annotations include masks for locally manipulated regions and human-annotated masks for structural or logical artifacts in fully AI-generated images, which serve as supervision for anomaly localization. To maintain a balanced data distribution, we use equal numbers of real and forged images and keep the sample counts across different forgery types approximately balanced.



Prompt of forged images:

You are a visual forensic assistant. Assess image authenticity using visible evidence. You are given two images of the same non-authentic sample. The first image is the original image. Its forgery type is {forgery_type} (locally manipulated / generatively edited / fully AI-generated / ...). The second image is an auxiliary evidence view constructed from the localization mask. Ordinary regions are darkened, while suspicious regions are enclosed by green boxes. Use this view only to locate areas that require closer inspection.

Analyze the image as follows:
1. Briefly describe the scene, main objects, and their relationships.
2. Inspect the suspicious regions and identify the most relevant visible evidence, such as:
   - unnatural boundaries or imperfect blending;
   - texture, noise, color, or illumination inconsistencies;
   - malformed anatomy, distorted geometry, or incomplete objects;
   - repeated textures, distorted text, or implausible spatial relationships;
   - inconsistent shadows, reflections, or perspective.
3. Explain why the observed evidence supports the specified forgery type.
4. End with a clear authenticity judgment.

Output requirements:
- Discuss only evidence visible in the original image.
- Refer to suspicious areas by semantic location, such as "the face" or "the background behind the object."
- Do not mention the auxiliary image, mask, green boxes, artifact description, ground-truth label, or annotation process.

Prompt of real images:

You are a visual forensic assistant. Assess image authenticity using visible evidence. You are given one authentic sample. The provided image is an authentic and unmodified photograph.

Analyze the image as follows:
1. Briefly describe the scene, main objects, and their relationships.
2. Inspect the image and identify the most relevant visible evidence of consistency, such as:
   - plausible object shapes, human anatomy, and spatial arrangements;
   - continuous textures, noise patterns, and colors;
   - consistent illumination, shadows, and reflections;
   - correct geometry and perspective;
   - reasonable physical and environmental relationships.
3. Explain why the observed consistency supports the conclusion that the image is natural and unmodified.
4. End with a clear authenticity judgment.

Output requirements:
- Discuss only evidence visible in the provided image.
- Focus on positive evidence of consistency and do not invent suspicious artifacts.
- Do not make unsupported assumptions about the image source or capture device.

Figure 5: Prompt templates used for data construction. The forged-image prompt directs the model to examine suspicious regions and explain the observed inconsistencies, whereas the real-image prompt guides it to assess overall visual and physical consistency.

Step 2: Generate the text analysis. We use Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025) to generate an authenticity analysis for each image. For forged samples, we create an auxiliary evidence view by darkening the unmarked background and outlining masked regions with green bounding boxes. This view preserves the content and spatial context of suspicious regions and directs the model's attention to them. Qwen3-VL receives both the original image and the auxiliary view for forged samples, and only the original image for real samples.

As shown in Figure 5, separate prompt templates are used for forged and real images. The forged-image prompt specifies the forgery type and asks the model to explain visible boundary, texture, lighting, structural, or spatial inconsistencies before making an authenticity judgment. The real-image prompt instead checks whether the scene is visually and physically consistent. All responses

must rely only on evidence visible in the original image and must not mention masks, bounding boxes, auxiliary views, or the annotation process. The resulting analysis can therefore be used directly as text supervision. We then manually prepend an authenticity judgment, either *real* or *forgery*, to each generated analysis.

Step 3: Manual quality control. Following prior data-curation practices (Wen et al., 2025; Kang et al., 2025), we apply a multi-stage quality-control procedure. We first perform rule-based checks to ensure the prepended authenticity judgment agrees with the source label, the response follows the required format, and it contains no references to masks, bounding boxes, or auxiliary views. We then verify the analysis against the original images and annotations to identify apparent mismatches. For forged images, the reported anomalies are checked for consistency with the visible content and localized regions; for real images, the analysis is checked for unsupported suspicious evidence. Samples that fail these checks are regenerated, while ambiguous cases are manually reviewed. We retain samples that satisfy these checks.

A.1.2 Training Configuration

Our Progressive Evidence Training consists of two stages: *Cross-Evidence Gated Alignment* and *Capability Refinement*. In Stage 1, the scheduled gate gradually increases the strength of cross-stream interaction, allowing the fine-grained and logical streams to progressively adapt to and integrate information from each other. Stage 2 further strengthens the model through two output-specific refinement steps, language-specialized training and localization-specialized training, thereby enhancing its image forgery detection capability. Table 7 summarizes the settings for all training runs. The complete training process takes approximately three days on eight NVIDIA RTX 4090 GPUs.

Table 7: Detailed configurations and loss weights for Progressive Evidence Training. The table outlines the specific hyperparameter settings used throughout training, including the trainable parameters, dataset scale, loss weights, and optimization schedule. These configurations reflect the transition from joint cross-evidence gated alignment to capability refinement of language and localization capabilities.

Stage	Setting	Trainable parameters	Training data	λ_{txt}	λ_{seg}	Epochs	Warmup	Learning rate	LoRA rank
Stage 1: Cross-Evidence Gated Alignment	LoRA	$\mathcal{P}_{F\to L}$, $\mathcal{P}_{L\to F}$, fine-grained expert layers, logical expert layers, and the fine-grained expert decoder	110K samples with masks and analysis	1.0	1.0	3	–	1.0×10^{-4}	512
Stage 2: Language-Specialized Training	LoRA	$\mathcal{P}_{F\to L}$, $\mathcal{P}_{L\to F}$, and logical expert layers	70K samples with analysis	1.0	0.0	1	0.05	1.0×10^{-5}	512
Stage 2: Localization-Specialized Training	SFT	$\mathcal{P}_{F\to L}$, $\mathcal{P}_{L\to F}$, fine-grained expert layers, and the fine-grained expert decoder	70K samples with masks	0.0	1.0	1	0.05	1.0×10^{-5}	–

Stage 1: Cross-Evidence Gated Alignment. Stage 1 uses the 110K-sample dataset constructed in the previous section, with each sample providing both an evidence localization mask and an authenticity explanation. During this stage, we fully fine-tune the two bridging projectors $\mathcal{P}_{\mathrm{F}\to\mathrm{L}}$ and $\mathcal{P}_{\mathrm{L}\to\mathrm{F}}$, together with the fine-grained expert decoder, while applying LoRA to the Fine-Grained Expert and Logical Expert layers. All other model components remain frozen.

The cross-evidence gate controls the attention assigned to bridged features at each BES layer. At the beginning of training, the gate suppresses the bridged-key logits while leaving the native-key logits unchanged, allowing each stream to rely primarily on its pretrained representations. Following a cosine schedule, the contribution of cross-stream information gradually increases to full strength during the first 10% of Stage 1 and remains fully enabled thereafter. We set $\epsilon = 10^{-6}$ for numerical stability and $T_w = 0.1T_1$, where T_w is the gate warm-up length and T_1 is the total number of Stage 1 training steps.

Stage 2: Capability Refinement. After cross-evidence alignment, Stage 2 sequentially refines the model's two output capabilities. It first strengthens language-based authenticity judgment and analysis, and then improves evidence localization.

Language-specialized training. We first use 70K samples from FakeClue (Wen et al., 2025), which contain a broader range of AI-generated images and expose the model to more diverse synthetic artifacts. The two cross-stream projectors and the logical expert layers are optimized with LoRA (Hu et al., 2022), while all other modules remain frozen. This step improves authenticity judgment and the generation of coherent, evidence-grounded analyses.

Localization-specialized training. We then select 70K samples with localization masks from the Stage 1 data. The two cross-stream projectors, the fine-grained expert layers, and the fine-grained expert decoder are optimized with supervised fine-tuning, while the language-related modules remain frozen. This step improves the accuracy and spatial precision of evidence localization without disrupting the language capability refined in the preceding step.

A.2 Additional Experiments

Layer Pairing in Hybrid Layer Interaction. Table 8 specifies the exact layer pairs used by Hybrid Layer Interaction. The fine-grained stream contains $N_{\mathrm{F}} = 32$ layers, and the logical stream contains $N_{\mathrm{L}} = 36$ layers. Following Eqs. (6) and (7) in the main paper, every third logical layer is selected from index 2 and paired with a fine-grained layer at the corresponding relative depth. The resulting 12 interaction pairs span the two streams at multiple depths, while all remaining layers continue along their stream-specific paths. All indices are zero-based. This arrangement enables recurring BES exchanges throughout the network while allowing the intervening layers to preserve stream-specific processing, thereby supporting sufficient cross-stream collaboration without excessive coupling.

Table 8: Specific layer pairs between the Fine-Grained Expert and Logical Expert used in Hybrid Layer Interaction. The two experts interact through 12 matched layer pairs, with zero-based layer indices.

Interaction pair	1	2	3	4	5	6	7	8	9	10	11	12
fine-grained expert layer	1	4	7	9	12	15	17	20	23	25	28	31
logical expert layer	2	5	8	11	14	17	20	23	26	29	32	35

Ablation of Cross-Evidence Gated Alignment. To further assess the effectiveness of Gated Alignment in Stage 1 of Progressive Evidence Training, we analyze its training dynamics by comparing the proposed cosine schedule, which gradually increases the strength of BES interaction, with a setting in which the gate is fully activated from the beginning. As shown in Figure 6, both settings reduce the loss rapidly at first, but the gated variant achieves a lower loss with smaller fluctuations later in training. This suggests that directly coupling the differently pretrained streams may introduce early cross-stream interference, whereas gradual activation allows the bridging projectors and expert layers to adapt progressively to incoming information from the other stream. Consequently, the two streams can integrate evidence more effectively and achieve better coordination, supporting the motivation of Progressive Evidence Training. This process establishes an initial alignment between the two streams, providing a foundation for subsequent capability refinement.

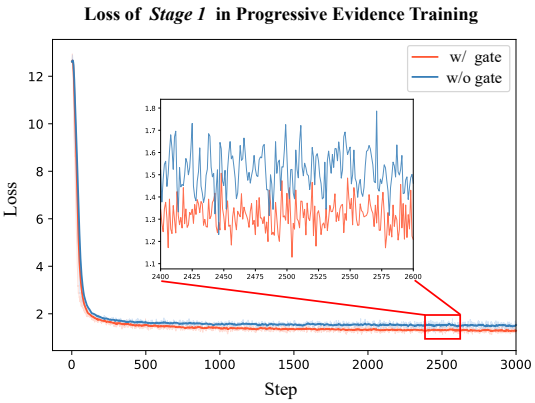


Figure 6: Training loss during Stage 1 of Progressive Evidence Training. The orange curve corresponds to the proposed cosine-scheduled gate activation, whereas the blue curve represents full gate activation from the start of training.

Robustness Experiments. We conduct robustness evaluations on CASIAv1+ (Dong et al., 2023), SynthScars (Kang et al., 2025), and COCO 2017 (Lin et al., 2014), representing locally manipulated, fully AI-generated, and real images, respectively. We evaluate TRACER under four common image perturbations: JPEG compression, Gaussian noise, Gaussian blur, and gamma correction. As shown in Table 9, performance is generally better preserved under JPEG compression and gamma correction, while stronger Gaussian noise and blur introduce larger variations. Fully AI-generated images remain comparatively stable under most settings, whereas locally manipulated and real images are somewhat more sensitive to certain perturbations. This difference may arise because the analysis of the latter two image types relies more heavily on fine-grained natural-image characteristics, which can be altered by common image processing. Overall, TRACER demonstrates good robustness across the evaluated image types and perturbations.

Table 9: Robustness analysis against common image distortions. We report mIoU and Acc to evaluate evidence localization and authenticity judgment capabilities, respectively, under varying levels of JPEG compression, Gaussian noise, Gaussian blur, and gamma correction.

Perturbation	CASIAv1+		SynthScars		COCO
	mIoU	Acc	mIoU	Acc	Acc
No perturbation	0.919	0.946	0.560	0.999	0.874
JPEG Comp. (QF = 90)	0.903	0.925	0.543	1.000	0.854
JPEG Comp. (QF = 80)	0.876	0.915	0.545	0.999	0.836
JPEG Comp. (QF = 70)	0.850	0.907	0.544	0.996	0.790
Gaussian Noise ($\sigma = 0.02$)	0.899	0.922	0.547	0.962	0.823
Gaussian Noise ($\sigma = 0.06$)	0.833	0.783	0.534	0.800	0.935
Gaussian Noise ($\sigma = 0.10$)	0.779	0.760	0.517	0.722	0.879
Gaussian Blur (Ksize = 3)	0.815	0.712	0.557	0.999	0.697
Gaussian Blur (Ksize = 5)	0.743	0.645	0.550	0.719	0.512
Gamma Correct. ($\gamma = 0.7$)	0.909	0.913	0.551	0.988	0.871
Gamma Correct. ($\gamma = 1.3$)	0.915	0.944	0.550	0.999	0.816

Balanced Judgment Accuracy Across Forgery forms. To assess whether authenticity judgments remain balanced when real and forged images are difficult to distinguish, we evaluate Columbia (Hsu & Chang, 2006), CocoGlide (Guillaro et al., 2023), and LOKI (Ye et al., 2025), representing traditional local manipulation, generative editing, and fully AI-generated images, respectively. Each benchmark contains forged and real images with broadly similar distributions. Table 10 reports accuracy for each type and their mean (bAcc). Several baselines favor real-image predictions, Qwen3.7 Max (Qwen Team, 2026) correctly detects 92.8% of real images but only 0.5% of forged images. TRACER achieves the highest bAcc among all compared methods on all three benchmarks. These results show that TRACER better balances forgery detection and real-image recognition across distinct forgery types.

Table 10: Balanced authenticity judgment across benchmarks representing diverse forgery forms, which include forged and real images with broadly similar distributions. $\mathrm{Acc}_{\mathrm{fake}}$ and $\mathrm{Acc}_{\mathrm{real}}$ are the class-wise accuracies, while bAcc is their equally weighted mean, $\frac{1}{2}(\mathrm{Acc}_{\mathrm{fake}} + \mathrm{Acc}_{\mathrm{real}})$. Higher is better. - indicates an unavailable result.

Method	Trad. Local Manipulation			Generative Edit			Fully AI-generated		
	Columbia			CocoGlide			LOKI		
	$\mathrm{Acc}_{\mathrm{fake}}$	$\mathrm{Acc}_{\mathrm{real}}$	bAcc	$\mathrm{Acc}_{\mathrm{fake}}$	$\mathrm{Acc}_{\mathrm{real}}$	bAcc	$\mathrm{Acc}_{\mathrm{fake}}$	$\mathrm{Acc}_{\mathrm{real}}$	bAcc
GPT-5.5 (OpenAI, 2026a)	.789	.891	.840	.006	.845	.426	.524	.734	.629
Gemini 3.1 Pro (Google DeepMind, 2026)	.902	.857	.880	.012	.661	.337	.576	.702	.639
Qwen3.7 Max (Qwen Team, 2026)	.839	.915	.877	.005	**.928**	.467	.654	**.897**	.776
FakeShield(22B) (Xu et al., 2025)	.448	.747	.598	**.843**	.059	.451	-	.409	-
FakeVLM(7B) (Wen et al., 2025)	-	.770	-	.027	.847	.437	.661	.774	.718
SIDA(13B) (Huang et al., 2025)	.655	**.931**	.793	.553	.894	.724	.057	.894	.476
TRACER(4B)	**.995**	.833	**.914**	.745	.762	**.754**	**.842**	.719	**.781**

False-Positive Localization on Real Images. Reliable anomaly evidence localization requires a model not only to identify forgery-related anomalous regions accurately but also to avoid producing spurious anomaly responses when the input is a real image. Therefore, We compare TRACER with several localization baselines on real images from COCO 2017 (Lin et al., 2014). Each predicted probability map is thresholded at 0.5. Pixel FPR measures the proportion of pixels incorrectly predicted as anomalous over the entire test set, while Image-FAR@0.5% measures the proportion of images whose predicted anomalous area exceeds 0.5% of the image. As shown in Table 11, TRACER achieves the lowest Pixel FPR and Image-FAR@0.5% among all compared methods. Together with its strong localization performance on forged images, these results show that TRACER can accurately localize anomaly evidence in forged content while suppressing erroneous localization on real images, demonstrating its reliable anomaly evidence identification capability.

Table 11: False-positive localization on real images from COCO 2017. Pixel FPR measures the proportion of pixels incorrectly identified as anomalous, while Image-FAR@0.5% measures the proportion of images whose predicted anomalous area exceeds 0.5% of the image. Lower values indicate better performance for both metrics.

Model	Pixel FPR	Image-FAR@0.5%
CAT-Net v2 (Kwon et al., 2022)	1.0304%	15.30%
Mesorch (Zhu et al., 2025)	0.7577%	18.64%
MVSS-Net (Dong et al., 2023)	1.6236%	23.22%
TruFor (Guillaro et al., 2023)	3.3373%	70.52%
SAFIRE (Kwon et al., 2025)	20.9546%	93.42%
SIDA (Huang et al., 2025)	1.3167%	13.48%
FakeShield (Xu et al., 2025)	11.0985%	76.78%
LEGION (Kang et al., 2025)	2.5545%	54.82%
TRACER	**0.2827%**	**9.68%**

A.3 More Visualization Examples

We provide additional qualitative examples to illustrate TRACER's anomaly evidence localization and complete image detection capabilities.

Dual-Assessment in Ambiguous Cases. Figure 7 shows that real images can contain capture-induced artifacts such as motion blur, while synthetic images may have no obvious perceptual anomalies. These cases illustrate the limitations of relying on a single evidence view and motivate TRACER's collaborative modeling of fine-grained and logical evidence, allowing visual details to be assessed alongside scene context. The model provides two output forms, an evidence mask at the visual level and a textual analysis at the language level, which jointly support authenticity assessment. In particular, an empty or weak mask does not establish authenticity or rule out forgery-related cues, which may be subtle, spatially diffuse, or difficult to delineate as a region. The textual analysis provides a complementary way to discuss these cues and to interpret unusual appearances in context. Together, these outputs provide users with a more comprehensive basis for assessing authenticity, including cases where conspicuous local anomalies are absent or potentially misleading.

Comparison of Evidence Localization Methods. Figure 8 presents representative cases comparing TRACER with TruFor (Guillaro et al., 2023), Mesorch (Zhu et al., 2025), FakeShield (Xu et al., 2025), LEGION (Kang et al., 2025), and SIDA (Huang et al., 2025) across diverse forgery scenarios. Across these forgery forms, TRACER consistently localizes anomaly evidence more accurately and produces masks that align more closely with the ground truth, while competing methods exhibit more missed or incorrectly localized regions. For example, in the generative-editing case involving two suitcases, TRACER accurately localizes both edited objects with well-aligned boundaries, whereas most competing methods either miss one of the suitcases or produce incomplete and noisy masks. These results demonstrate the clear advantage of TRACER in unified anomaly evidence localization.

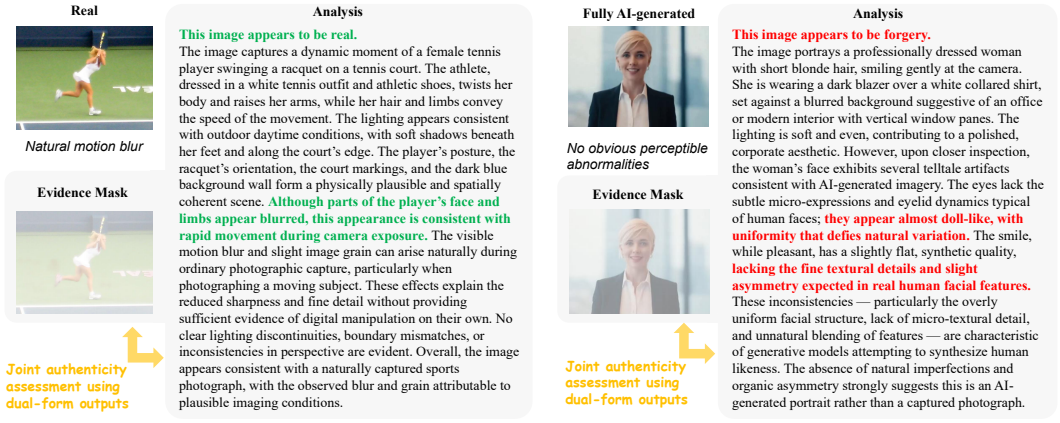


Figure 7: Ambiguous cases for visual authenticity assessment. Left: a real image with natural motion blur. Right: a fully AI-generated portrait without obvious perceptual anomalies.

Image Forgery Detection Capabilities. Figures 9 and 10 demonstrate TRACER's comprehensive image forgery detection capabilities across diverse forgery scenarios and authentic-image settings. For each input, TRACER jointly provides an anomaly evidence mask, an authenticity judgment, and an evidence-grounded analysis. For example, in the locally manipulated milk-carton image, TRACER accurately localizes the tilted carton and associates its forgery judgment with the object's implausible orientation and visual inconsistencies. In the fully AI-generated portrait, TRACER correctly identifies the image as forged, localizes the malformed fingers, and attributes its judgment to the corresponding anatomical inconsistencies. Together, these complementary forms of evidence enable TRACER to assess image authenticity from multiple perspectives, supporting comprehensive and practical forensic detection.

MLLM-as-a-Judge Evaluation of Explanation Quality. Figure 11 shows the system prompts used to calculate the Explanation Quality Score (EQS) in Table 2 of the main paper. We use Claude Sonnet 5 (Anthropic, 2026) as the judge. For forged images, the model receives the original image, the ground-truth forgery category and aligned evidence mask, and the candidate explanation. And for real images, it receives the original image and candidate explanation without a mask. Separate prompts score visual evidence, forensic reasoning, spatial grounding, and calibration, with the four subscores summed to a 0–10 EQS for a comprehensive evaluation of anomaly analysis capability.

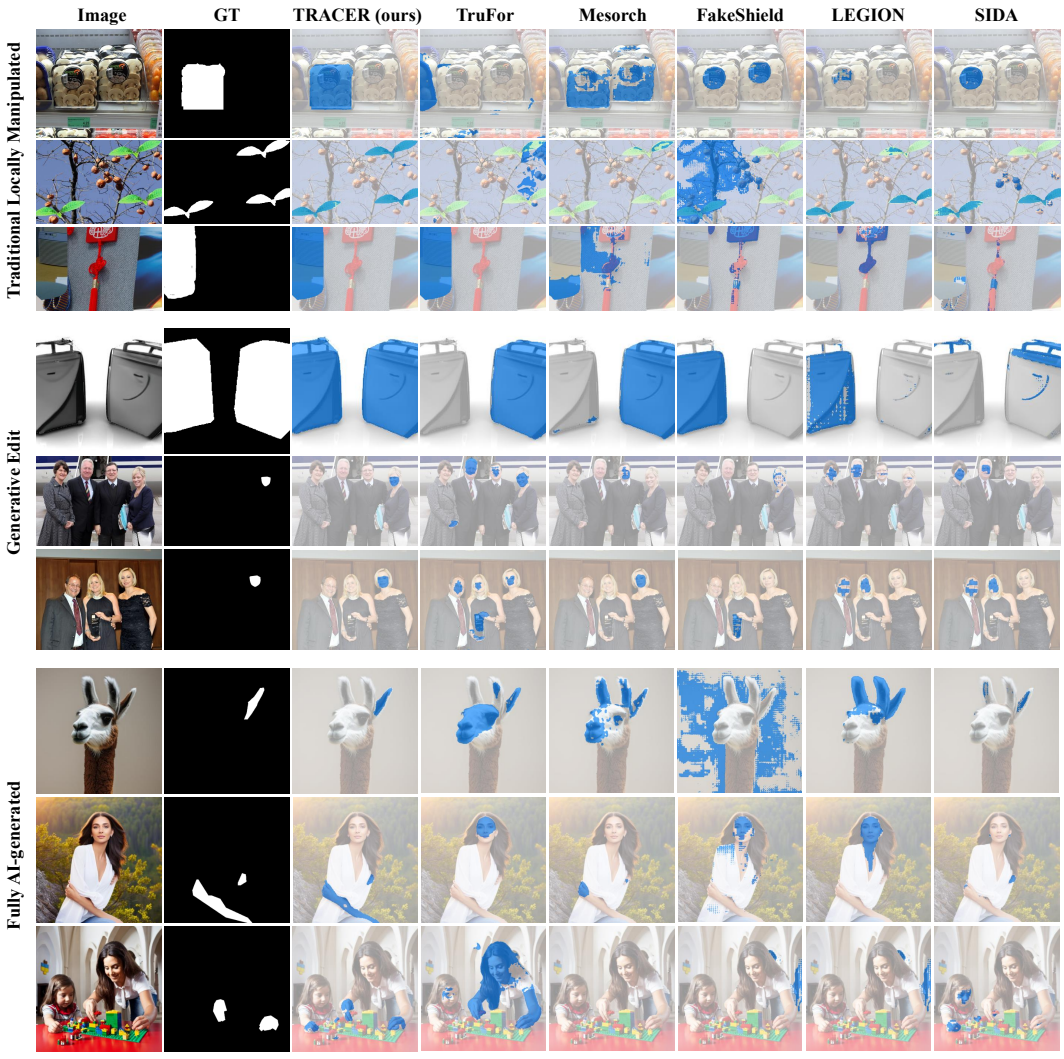


Figure 8: Qualitative comparison of anomaly evidence localization across traditional local manipulation, generative editing, and fully AI-generated images. All compared models are fine-tuned on the same training data as TRACER.

Traditional Locally Manipulated

Evidence Mask



Analysis

This image appears to be forged.
The image shows a supermarket shelf stocked with Dutch Lady full cream milk cartons, **the one on the right is visibly tilted at an unnatural angle**, suggesting it was digitally manipulated. The tampered region, **showing signs of digital manipulation**, encompasses the tilted Dutch Lady milk carton. The primary forgery trace is the severe distortion of the carton's orientation — it leans diagonally across the shelf, defying gravity and physical plausibility. The edges of the carton within the tampered area exhibit slight misalignment with the shelf's surface and adjacent products, indicating it was likely pasted or rotated into place. Additionally, the lighting and shadowing on the tilted carton do not match the ambient illumination of the surrounding items; the highlights and shadows appear inconsistent with the direction and intensity of light in the scene. The background behind the tilted carton also lacks the natural depth and perspective continuity seen in the upright cartons, further suggesting digital insertion. These artifacts collectively point to deliberate manipulation rather than a naturally occurring object.

Traditional Locally Manipulated

Evidence Mask



Analysis

This image appears to be forged.
The image depicts an ancient Egyptian temple complex with towering columns and scattered ruins, but **a large tortoise has been digitally inserted into the foreground**, creating a surreal and incongruous scene. **The tortoise is clearly a forged element, evidenced by several visual inconsistencies.** First, its lighting and shadows do not match the ambient illumination of the temple: the tortoise appears evenly lit with no corresponding shadow cast on the ground beneath it, while surrounding structures exhibit directional sunlight and subtle shading. Second, the texture of the tortoise's shell and skin lacks the fine detail and natural wear expected in a real-world photograph; it appears overly smooth and slightly blurred, inconsistent with the crispness of nearby architectural elements. Third, the edges of the tortoise are unnaturally sharp and well-defined against the background, suggesting digital compositing rather than natural occlusion or depth-of-field blur. Finally, the scale and perspective of the tortoise relative to the temple structures are implausible — it appears disproportionately large and positioned as if floating or walking on the ground without any supporting context or interaction with the environment. In conclusion, the tortoise is a digitally added forgery, identifiable through mismatched lighting, texture anomalies, edge artifacts, and spatial inconsistencies that betray its artificial insertion into the scene.

Generative Edit

Evidence Mask



Analysis

This image appears to be forged.
A freshly baked pizza with a vibrant red sauce and green toppings sits on a metal tray, accompanied by a small glass jar of red condiment, all set against a dimly lit, warm-toned background. **The tampered region, showing signs of digital manipulation, encompasses the entire pizza.** The edges of the pizza within this area exhibit unnatural sharpness and slight haloing against the darker background, suggesting that it is not an original component of the captured scene. The lighting on the pizza is inconsistent with the ambient illumination of the surrounding table and background — the pizza appears overly bright and evenly lit, while the background has softer, directional shadows that do not match the highlights on the food. Additionally, the texture of the crust and sauce lacks the subtle imperfections and depth expected from a real photograph, appearing slightly smudged or airbrushed. These anomalies suggest the pizza is **plausibly an AI-generated element blended into the frame**, instead of being recorded alongside the surroundings during the original capture. Conclusion: The pizza is a forged element, artificially inserted into the scene with visible edge artifacts and lighting mismatches, confirming digital manipulation.

Generative Edit

Evidence Mask



Analysis

This image appears to be forged.
In a crowded indoor event, four men are engaged in conversation, with one man's face showing signs of digital manipulation. **The forged region is centered on the face of the man in the middle, who is wearing a black shirt and has sunglasses hanging from his collar.** The tampering is evident through several artifacts: the skin texture around the eyes and forehead appears unnaturally smooth and lacks the fine wrinkles and pores visible on the surrounding faces; there is a subtle halo or edge discontinuity along the jawline and cheek where the face meets the shirt collar, suggesting a cut-and-paste or generative synthesis; lighting on the face is inconsistent with the ambient illumination of the scene — it appears slightly brighter and more evenly lit than adjacent faces, which show natural shadows and highlights from overhead lights. Additionally, the glasses hanging from the collar do not cast any corresponding shadow on the shirt, further indicating an artificial insertion. These inconsistencies point to **a deepfake or face-swapping manipulation targeting this individual.** Conclusion: The central face has been digitally altered, exhibiting clear signs of forgery through texture smoothing, edge artifacts, and lighting mismatches.

Figure 9: Complete outputs from TRACER for traditional local manipulations and generative edits, including predicted evidence masks, authenticity judgments, and corresponding analyses.

Fully AI-generated | Evidence Mask




Analysis

This image appears to be forged.
This black-and-white portrait shows a woman with dark, shoulder-length hair gazing directly at the viewer. Her hand is raised to her chin, with her fingers gently curled, and she wears a ring on her ring finger. Soft, out-of-focus lights in the background suggest a studio or nighttime setting. **Several inconsistencies suggest that the image is AI-generated**. Most noticeably, **the fingers appear fused and misshapen, lacking the clear definition and joint structure of a real human hand**. The thumb seems to merge into the index finger, creating an anatomically impossible form. The ring finger also has a malformed nail bed, with the nail appearing to grow from the side rather than the tip. Its irregular shape is inconsistent with natural nail growth. In addition, the lighting on the hand does not fully match the sharper highlights and shadows on the face and hair, while the shadows beneath the fingers appear overly soft or misplaced. These anatomical distortions, unnatural textures, and lighting mismatches are common artifacts in AI-generated images. Although the overall portrait appears realistic, these small structural errors strongly suggest that it was artificially synthesized.

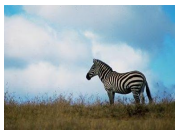
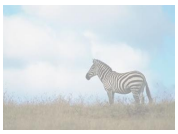
Fully AI-generated | Evidence Mask




Analysis

This image appears to be forged.
The image portrays a tiger resting on a stone surface, surrounded by dense green foliage. Its relaxed posture and direct gaze suggest a calm but alert state, with its body facing the viewer and its head slightly turned. However, **several anatomical and textural inconsistencies suggest that the image is AI-generated**. The most noticeable anomaly is the **tiger's left front paw**, which appears unnaturally elongated and misshapen. Its toes seem fused or partly missing, lacking the clear separation and structure expected in a real feline limb. The fur around the paw and lower foreleg also has a smudged, painted quality rather than the fine, directional strands of real fur. In addition, the connection between the paw and the stone surface is unclear. The paw appears to float slightly above the ground, without convincing contact shadows or signs of pressure against the surface. The lighting in this area also seems inconsistent with that on the tiger's face and body. These anatomical distortions, unnatural textures, and lighting inconsistencies are common artifacts in AI-generated images. Although the overall scene appears realistic, these structural errors strongly suggest that it was synthesized rather than captured by a camera.

Real | Evidence Mask

Analysis

This image appears to be real.
The image captures a solitary zebra standing in profile on a grassy savanna under a partly cloudy sky, presenting a natural wildlife scene. The zebra is positioned slightly off-center, facing left, with its body aligned parallel to the frame, allowing for a clear view of its striped pattern and muscular build. The lighting appears consistent with natural daylight—soft shadows fall beneath the zebra and along the terrain, suggesting an overcast or diffused light source, while highlights on the zebra's coat and the grassy foreground indicate ambient illumination from above. The background features a layered sky with soft cumulus clouds, which recede into atmospheric haze, creating a sense of depth that aligns with realistic perspective. The zebra's posture, the texture of the dry grass, and the subtle motion blur in the distant vegetation all contribute to a dynamic yet serene moment. **There are no visible signs of digital manipulation**: edges are naturally defined, lighting is coherent across all elements, and the environmental context (open plains, sky, vegetation) is ecologically plausible. This is a genuine, unedited photograph capturing a real moment in nature.

Real | Evidence Mask




Analysis

This image appears to be real.
The image captures a serene riverside scene featuring the historic City Palace complex in Udaipur, India, bathed in the warm, golden light of late afternoon or early evening. The palace's expansive stone architecture, with its tiered structures and ornate balconies, dominates the midground, rising majestically above lush green trees that frame the lower portion of the composition. In the foreground, calm waters reflect the ambient glow, with a small red boat visible on the right, adding scale and life to the tranquil setting. A dark silhouette of overhanging leaves occupies the upper left corner, creating a natural vignette that enhances depth and perspective. The lighting is consistent with a low sun angle, casting soft shadows and highlighting textures without harsh contrasts; the water's gentle ripples and the subtle haze in the sky suggest atmospheric conditions typical of a warm, humid day. All elements — from the architectural alignment and structural details to the natural reflections and foliage — **adhere to physical laws and environmental logic, with no visible signs of digital manipulation or artificial compositing**. The image appears to be a genuine, unedited photograph captured in natural conditions.

Figure 10: Complete outputs from TRACER for fully AI-generated and real images, including predicted evidence masks, authenticity judgments, and corresponding analyses.

MLLM-as-a-Judge System Prompts

 # Forgery

You are an image forensics expert. Evaluate a candidate explanation for a forged image. The ground-truth category and mask meaning are given in the user message.

Inspect the original image and use the aligned mask to check the explanation. The mask may be incomplete; clearly visible evidence outside it also counts. Do not require the candidate to mention the mask or name the exact forgery type.

Award integer points:
1. Visual evidence (0–4): Are the described anomalies specific, visible, and accurate?
2. Forensic reasoning (0–3): Does the explanation show why those observations indicate that the image is inauthentic? A "fake" label alone earns no points here.
3. Spatial grounding (0–2): Does it accurately describe where the relevant anomalies appear?
4. Calibration (0–1): Does it avoid unsupported certainty and stay focused on visible evidence?

Do not reward length or the number of claimed anomalies. Score only accurate, relevant, visually supported claims. Do not penalize the candidate for omitting a masked artifact that is not visually discernible. Sum the four subscores for a total of 0–10.

Return only:

```
{
 "visual_evidence": <0-4>,
  "forensic_reasoning": <0-3>,
  "spatial_grounding": <0-2>,
  "calibration": <0-1>,
  "score": <sum of the four subscores>,
  "reason": "<brief image-grounded justification>"
}
```

 # Real

You are an image forensics judge. The ground-truth category is Real. The user message contains the original image followed by a candidate explanation. No mask is provided.

Inspect the image and award integer points:
1. Visual evidence (0–4): Are the described details accurate, without invented forgery artifacts?
2. Forensic reasoning (0–3): Does the explanation justify why no convincing forgery evidence is visible? A "real" label alone earns no points here.
3. Spatial grounding (0–2): Are its observations tied to actual image content? Do not require an anomaly location.
4. Calibration (0–1): Does it avoid claiming that visual inspection proves authenticity?

Do not reward length or require the candidate to invent positive proof that the image is real. Score each criterion separately and sum the points for a total of 0–10.

Return only:

```
{
 "visual_evidence": <integer 0-4>,
  "forensic_reasoning": <integer 0-3>,
  "spatial_grounding": <integer 0-2>,
  "calibration": <integer 0-1>,
  "score": <sum of the four subscores>,
  "reason": "<brief image-grounded
justification>"
}
```

Figure 11: System prompts for evaluating Explanation Quality Score (EQS) with an MLLM judge.