EVOLVING MULTI-MODAL LARGE LANGUAGE MODEL WITHOUT TRAINING FOR ZERO-SHOT IMAGE FORGERY LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Image Forgery Localization(IFL) aims to identify and precisely locate tampered regions in images. To tackle this practical but challenging task, existing methods typically locate forged region in different scenarios by employing a large amount of pre-collected training data. However, with the rapid iteration and upgrading of forgery strategies, new forgery localization requirements are often accompanied by training data that cannot be collected in advance, making it difficult for existing methods to fail against the latest forgery techniques. In this article, we focus on a such practical and challenging task, namely Zero-shot Image Forgery Localization(ZIFL). Inspired by recent advances in Multi-modal Large Language Models(MLLMs), we propose a training-free image forgery localization method called **C**ollaborative **LEA**der-worker discrimi**N**ator, called **CLEAN**, which performs zero-shot forgery localization by gathering the cross-modal reasoning capacity of MLLMs and the fine-grained localization capacity of Segment Anything(SAM) to identify forgery clues. Specifically, we propose a Coarse-Grained Suspect Exploration(CGSE) module to act as a forensic analysis leader and systematically generate a list of candidate suspect regions based on high-level forgery evidence. Besides, a Fine-Grained Iterative Refinement(FGIR) module is designed to judge the forgery clues in an iterative manner, which eliminates normal backgrounds and refines the boundaries of tampered regions for accurate ZIFL. Extensive experiments on multiple benchmarks demonstrate that our method achieves competitive IFL performance against existing training-based methods without relying on pre-collected training data on zero-shot image forgery localization task. **Our code will be released soon.**

1 INTRODUCTION

Image Forgery Localization (IFL) aims to locate manipulated regions in images at the pixel level. To this end, existing methods typically rely on training on massive forged images (Dong et al., 2022; Gao et al., 2022) to develop the capacity of forgery traces localization, which are distributed in the spatial (Bappy et al., 2019; Gao et al., 2022) and frequency domains (Kwon et al., 2022; Rao et al., 2022). However, as shown in Figure 1(a), with the rapid evolution of forgery technologies, new forgery methods emerge in an unpredictable way, where insufficient training data often leads to a decline in the practicality due to insufficient training timeliness. Therefore, this paper focuses on a practical and challenging task, called Zero-shot Image Forgery Localization(ZIFL), aiming to locate forged content manipulated by different methods without relying on any training images.

To tackle the above problem, the Multi-modal Large Language Models (MLLMs) (Team, 2025; Yang et al., 2025; Liu et al., 2023a; Hong et al., 2025) have become a feasible solution, which has demonstrated remarkable capabilities in cross-modal reasoning, commonsense understanding, and structured scene interpretation across diverse vision-language tasks, including image retrieval (Park & Lee, 2025), semantic segmentation (Zhang et al., 2023; Wang et al., 2023), and multi-modal understanding (Lee et al., 2025; Wang et al., 2025), as well as zero-shot prediction (Kojima et al., 2022; Naeem et al., 2023; Li et al., 2025). However, it is extremely challenging to transfer pre-trained large-scale content recognition capabilities to forgery clue detection for the ZIFL task when it comes to alternative recognition of semantic plausibility, physical consistency, and contextual coherence in a

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.

training-free manner. Fortunately, we observe that forgery detection can be achieved by transforming the detection of forgery traces into visual semantic information that violates prior knowledge involved in MLLMs, thereby comparing the inconsistencies between information in the image and prior information, rather than detecting forgery as the identification of manipulation-specific artifacts. This perspective enables a functional transfer: instead of assuming that MLLMs possess explicit forensic knowledge, we leverage their implicit prior-modeling capability as a prior-violation analyzer. In other words, we migrate the strong multi-modal reasoning capacity of MLLMs to the forgery localization task by repurposing their consistency reasoning ability to identify anomaly hypotheses in suspicious regions. However, MLLMs alone lack precise spatial grounding and pixel-level discrimination. A naive combination with segmentation models (e.g., SAM) (Kirillov et al., 2023; Ravi et al., 2024) is insufficient, as it does not explicitly model how reasoning outputs should guide localization under complex and multi-region manipulation scenarios. Therefore, an effective mechanism is required to align high-level anomaly reasoning with fine-grained spatial refinement.

To address the above challenges in the ZIFL, we propose a training-free Collaborative LEAder-worker discrimiNator (CLEAN) framework, which aims to locate forged regions without training. Our core idea is to establish a reasoning-guided localization paradigm that explicitly transfers the multi-modal reasoning capability of MLLMs to pixel-level forgery localization. Specifically, we decompose the task into two synergistic stages: anomaly hypothesis generation and iterative spatial refinement. In the first stage, a Coarse-Grained Suspect Exploration (CGSE) module is proposed to act as a forensic analyze leader to systematically generate a list of candidate suspect regions based on high-level forgery evidence. In the second stage, a Fine-Grained Iterative Refinement (FGIR) module is introduced to act as a refinement localization worker to distinguish discrepancies between different types of forged clues from different perspectives in an iterative manner. By explicitly modeling the interaction between multi-modal consistency reasoning and fine-grained spatial grounding, CLEAN effectively bridges high-level anomaly reasoning with precise pixel-level localization, leading to strong generalization to unseen forgery strategies without additional training.

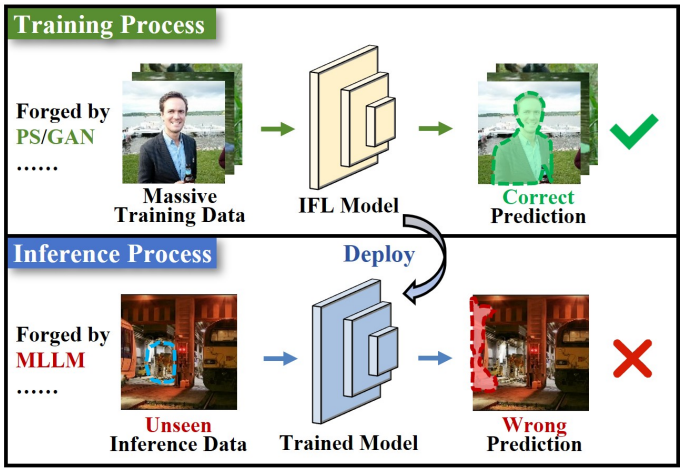


Figure 1: The influence of the rapid evolution of forgery technologies, where the trained IFL model have difficulties in applying to real-world deployments because the forgery methods encountered during testing (*e.g.*, forgery by Multi-modal Large Language Models(MLLM) (Xu et al., 2024)) differ from those seen during training (*e.g.*, forgery by Photoshop(PS) (Dong et al., 2013) and Generative Adversarial Networks (GAN) (Novozamsky et al., 2020)).

In summary, the main contributions of this paper are as follows: (1) A training-free Collaborative LEAder-worker discrimiNator (CLEAN) framework is proposed to tackle the Zero-shot Image Forgery Localization(ZIFL) task, which generates pixel-level forgery results by gathering the multi-modal consistency reasoning capacity of MLLMs and the fine-grained localization capacity of SAM to identify forgery clues. (2) A Coarse-Grained Suspect Exploration (CGSE) module is introduced to discover potential forgery regions in images based on different forgery clues by exploiting the reasoning potential of large models for forgery traces. (3) A Fine-Grained Iterative Refinement (FGIR) module is designed to judge the forgery clues in an iterative manner, which eliminates normal backgrounds and refines the boundaries of tampered regions for accurate zero-shot image forgery localization.

2 Related Work

2.1 Image Forgery Localization

Image Forgery Localization (IFL) aims to identify and locate tampered regions within an image. Early traditional methods (Ferrara et al., 2012; Carvalho et al., 2015; Mahdian & Saic, 2009) for addressing the IFL problem relied on hand-crafted features to capture statistical anomalies introduced by specific forgery types,such as Color Filter Array (CFA) patterns or illumination inconsistencies. However, due to their dependence on prior knowledge and limited generalization capability, these methods struggle to handle diverse forgery types in real-world scenarios. Recent deep learning based IFL methods mainly acquire task-specific knowledge in two ways. The first line follows a fully supervised training paradigm, where task-specific localization networks are trained on annotated forged images to learn discriminative tampering artifacts (Bappy et al., 2019; Bi et al., 2019; Dong et al., 2022; Hu et al., 2020; Li & Huang, 2019; Liu et al., 2023b). The second line adapts pre-trained or foundation models to forgery-related perception tasks through task-specific tuning or prompting strategies (Zhou et al., 2020; 2018; Zhuang et al., 2021). Despite their progress, both paradigms still rely on task-specific supervision or adaptation, limiting their flexibility under unseen forgery distributions. Therefore, in this paper, we focus on the Zero-shot Image Forgery Localization task. To enable the zero-shot prediction capacities, recent research has begun exploring the application of Multi-modal large language models (MLLMs).

2.2 Multi-modal Large Language Models

Multi-modal Large Language Models (MLLMs) are an emerging paradigm capable of cross-modal understanding with strong zero-shot generalization and semantic reasoning capabilities. Inspired by the success of MLLMs in visual tasks (Sun et al., 2025; Zhang et al., 2024; Lee et al., 2025; Li et al., 2025), several studies have begun to apply MLLMs to deepfake detection tasks. However, applying MLLMs to the pixel-level IFL task reveals significant limitations. Their outputs are typically high-level textual descriptions, lacking the precise pixel-level localization capability required by IFL. More importantly, a naive application of MLLMs is insufficient. They struggle to handle the subtle, fine-grained differences between forged and real information and often fail to capture varied forensic traces from a global perspective when multiple forgery types are present. Therefore, to achieve precise pixel-level localization, fine-grained detection models should be integrated to compensate for their limitations in spatial precision and fine-grained analysis.

2.3 Segment Anything Model

Segment Anything Model (SAM) (Kirillov et al., 2023) is a foundational segmentation model capable of high-quality image segmentation based on diverse prompts, with SAM2 (Ravi et al., 2024) and SAM3 (Carion et al., 2025) extending capabilities to video domains through streaming and spatiotemporal memory mechanisms. These models have inspired numerous extensions and pipelines. For example, LangSAM integrates language-conditioned object detection with SAM to enable text-guided segmentation, while HQ-SAM (Ke et al., 2023) improves segmentation quality through High-Quality Tokens. However, despite SAM's powerful segmentation capabilities, it lacks prior understanding of forgery clues, making it difficult to autonomously identify and localize forged regions without precise prompts.

Different from existing methods, we propose CLEAN, a zero-shot image forgery localization framework without any parameter updating, which achieves efficient Zero-shot Image Forgery Localization through coarse-to-fine suspect exploration and iterative refinement, thus avoiding the rely on training data.

3 Method

3.1 Preliminary

This paper focuses on the Zero-shot Image Forgery Localization (ZIFL) task, which aims to localize manipulated regions from unseen forgery patterns without task-specific training or parameter adaptation. Formally, let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a test set, where x_i is an input image and y_i is the

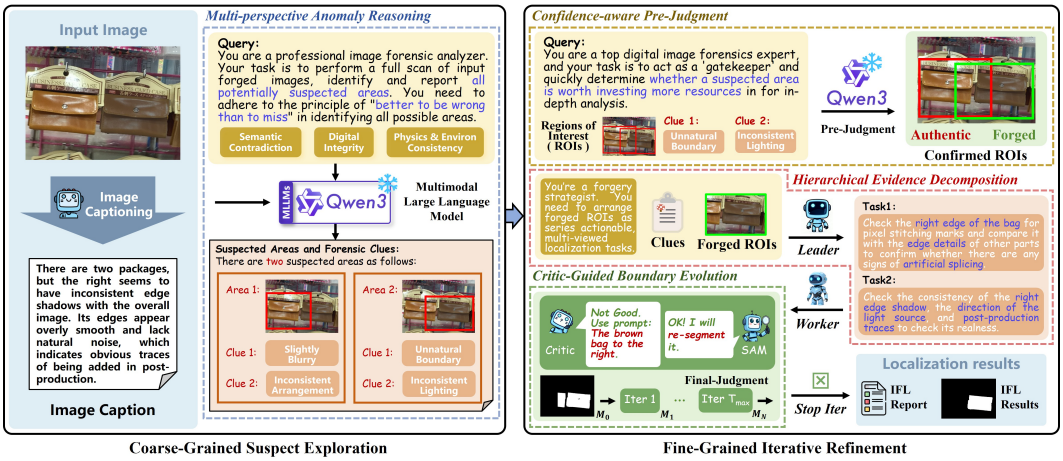


Figure 2: The overall architecture of Our CLEAN framework, which consists of a Coarse-Grained Suspect Exploration (CGSE) module and a Fine-Grained Iterative Refinement (FGIR) module. The CGSE module performs multi-perspective anomaly reasoning to preliminary locate suspected areas and generate forensic clues for each area, while the FGIR module confirms them with confidence-aware pre-judgment and introduces, and generate localization results through hierarchical evidence decomposition and critic-guided boundary evolution.

corresponding pixel-level forgery mask. Different from traditional training-based IFL methods, ZIFL does not update model parameters through task-specific training or adaptation on the target forgery distribution. Given an input image x_i, the goal of ZIFL is to directly predict a pixel-level localization mask $\hat{y}_i$, which is expected to accurately identify manipulated region under diverse and previously unseen forgery patterns.

3.2 Overview

As shown in Figure 2, the pipeline of our proposed CLEAN consists of a Coarse-Grained Suspect Exploration (CGSE) module and a Fine-Grained Iterative Refinement (FGIR) module. Given an input image x_i, the CGSE module first exploits the reasoning ability of MLLMs to discover suspected areas from multi-dimensional forensic clues. Then, the FGIR module filters false-positive candidates and refines the supported suspected areas into pixel-level forgery masks by progressively grounding verified forensic clues into segmentation masks.

3.3 Coarse-Grained Suspect Exploration

The **Coarse-Grained Suspect Exploration** (CGSE) module aims to retrieve high-recall suspect proposals before fine-grained localization. Since MLLMs are more capable of high-level forensic reasoning than precise pixel-level grounding, CGSE first establishes a global scene context and then explores potential anomalies from complementary forensic perspectives, rather than directly predicting the final forgery mask. As shown in the left part of Figure 2, this process transforms open-ended forensic reasoning into evidence-grounded suspect proposals for subsequent refinement.

Multi-Perspective Anomaly Reasoning. Given an input image x_i, we first obtain a global caption to provide a scene-level reference. Based on the input image and the caption, the MLLM acts as a forensic analyzer to conduct multi-perspective anomaly reasoning. Rather than relying on isolated visual cues, it scans the image through three complementary forensic dimensions: semantic contradiction, digital integrity, and physical-environment consistency. By jointly considering these three complementary dimensions, CGSE derives more reliable forensic clues for suspected regions and reduces suspicious judgments caused by isolated or ambiguous observations. The detailed prompt used for this process is provided in Appendix I.

To bridge open-ended reasoning with subsequent pixel-level localization, the MLLM's outputs must be formalized into a structured suspect proposal set:

$$\mathcal{P} = \{p_i\}_{i=1}^{N_p}, \quad p_i = (desc_i, bbox_i, clues_i, conf_i), \tag{1}$$

where each p_i encapsulates the textual description $desc_i$, the initial bounding box $bbox_i$, the forensic clues $clues_i$, and the confidence score $conf_i$. In this way, each proposal is not merely a coarse region candidate, but an evidence-grounded suspect proposal that indicates both where the suspicious area may appear and why it should be further verified. This structured representation bridges MLLM-based forensic reasoning and the following pixel-level spatial refinement.

However, the bounding boxes generated by MLLMs are usually coarse and object-centric. They may cover only the main suspicious object while omitting boundary artifacts or nearby contextual evidence that is important for forgery verification. To alleviate this localization bias and recover a sufficient forensic receptive field, we introduce a context-aware expansion mechanism dynamically adjusted by the relative scale of each proposal. For a proposal p_i bounded by $bbox_i = (u_i, v_i, w_i, h_i)$, its area ratio relative to the input image is computed as:

$$r_i = \frac{area(bbox_i)}{area(x_i)}. \tag{2}$$

The adaptive expansion ratio α_i is then determined by:

$$\alpha_i = \begin{cases} \alpha_h, & r_i < \tau_1, \\ \alpha_m, & \tau_1 \le r_i \le \tau_2, \\ \alpha_s, & r_i > \tau_2, \end{cases} \tag{3}$$

where smaller candidate regions are explicitly assigned larger expansion ratios to preserve a broader surrounding context. The context-aware bounding box is subsequently computed as:

$$bbox'_i = (u_i - \alpha_i w_i,\ v_i - \alpha_i h_i,\ (1 + 2\alpha_i)w_i,\ (1 + 2\alpha_i)h_i), \tag{4}$$

with coordinates rigorously clipped to the image boundaries. Ultimately, the expanded proposal set $\mathcal{P}' = \{p'_i\}_{i=1}^{N_p}$ provides context-rich candidate regions, serving as a reliable initialization for the hierarchical decomposition and verification in the FGIR module.

3.4 Fine-Grained Iterative Refinement

While the CGSE module successfully retrieves an expanded proposal set $\mathcal{P}'$, these coarse proposals may still contain false positives and exhibit loose spatial boundaries. For each expanded proposal p'_i, we crop the corresponding Region of Interest (ROI) from the input image according to its expanded bounding box $bbox'_i$. To further verify these candidate ROIs and refine their coarse localization, we propose the **Fine-Grained Iterative Refinement (FGIR)** module, which progressively filters unreliable regions and grounds the verified forensic clues into pixel-level forgery masks.

Confidence-aware Pre-Judgment. Since the candidate ROIs may include authentic regions caused by natural scene variations, FGIR first performs pre-judgment to filter unreliable ROIs before fine-grained segmentation. For each candidate ROI R_i, the MLLM first verifies whether the provided forensic clues are visually supported and whether the ROI deserves further refinement. ROIs with weak, ambiguous, or naturally explainable clues are filtered out. For the Confirmed ROIs, we compute a confidence-aware score by jointly considering the proposal confidence $conf_i$ and the richness of its forensic clues:

$$score_i = \lambda \cdot S_{\text{prior}}(conf_i) + (1 - \lambda) \cdot \frac{|clues_i|}{\max_j |clues_j|}, \tag{5}$$

where $S_{\text{prior}}(\cdot)$ maps $conf_i \in \{\text{High}, \text{Medium}, \text{Low}\}$ to $\{1.0, 0.5, 0.0\}$. $|clues_i|$ denotes the number of forensic clues inherited from p'_i, and λ balances proposal confidence and clue richness. The retained ROIs are sorted by $score_i$ and sequentially passed to the subsequent decomposition stage.

Hierarchical Evidence Decomposition. For each Forged ROI R_i, its forensic clues may still be abstract and entangled, making direct prompt construction unstable. To bridge this semantic-to-visual gap, FGIR employs a Leader-Worker mechanism to decompose coarse forensic descriptions into

fine-grained verifiable sub-tasks. Given R_i and $clues_i$, the Leader generates a set of evidence-specific sub-tasks:

$$\{T_j\}_{j=1}^{K_i} = \mathrm{MLLM}_{\mathrm{Leader}}(R_i, clues_i), \tag{6}$$

where each T_j corresponds to a specific forensic cue or evidence dimension. Each Worker then verifies one sub-task and outputs a verdict, an evidence description, and a segmentation-oriented prompt:

$$(v_j, e_j, q_j) = \mathrm{MLLM}_{\mathrm{Worker}_j}(R_i, T_j), \tag{7}$$

where $v_j \in \{\mathrm{Support}, \mathrm{Reject}\}$ indicates whether the clue is visually supported, e_j describes the supporting visual evidence, and q_j denotes the prompt derived from the verified clue. Only prompts associated with supported clues are aggregated:

$$Q_i^0 = \mathrm{Aggregate}\left(\{q_j \mid v_j = \mathrm{Support}\}\right). \tag{8}$$

If all clues are rejected, the ROI is filtered out; otherwise, Q_i^0 serves as the initial segmentation prompt for the following boundary evolution. This decomposition transforms abstract forensic suspicions into verified evidence and segmentation-oriented prompts, providing a reliable initialization for pixel-level refinement.

Critic-Guided Boundary Evolution. Although the supported prompt Q_i^0 provides a localization prior, a single-pass segmentation may still follow natural object contours rather than true tampered boundaries. Therefore, FGIR introduces critic-guided boundary evolution to refine the ROI mask according to forensic consistency. At iteration t, LangSAM generates a local mask from the current prompt:

$$M_i^t = \mathrm{LangSAM}(R_i, Q_i^t), \tag{9}$$

where Q_i^0 is initialized from the supported prompts produced by hierarchical evidence decomposition. The Critic then evaluates the current mask with respect to the Forged ROI and the verified evidence:

$$a_i^t = \mathrm{Critic}(R_i, M_i^t, \{e_j \mid v_j = \mathrm{Support}\}), \tag{10}$$

where a_i^t denotes the critic feedback. If the mask sufficiently covers the supported evidence while avoiding excessive normal background, the Critic returns a stop signal. Otherwise, the feedback is used to update the segmentation prompt:

$$Q_i^{t+1} = \mathrm{Refine}(Q_i^t, a_i^t). \tag{11}$$

Through this feedback loop, the mask boundary progressively expands toward missing suspicious evidence, shrinks away from authentic background, or shifts to better align with the manipulated region.

The refinement stops when the Critic considers the mask reliable or when two adjacent masks become stable:

$$\mathrm{IoU}(M_i^t, M_i^{t-1}) > \beta, \tag{12}$$

where β denotes the convergence threshold. If the maximum iteration number $T_{\max}$ is reached, the latest mask is retained.

Final-Judgment and Localization Results Generation. After boundary evolution, each retained forged ROI produces a refined local mask. To obtain the image-level localization result, the local mask is mapped back to the coordinate system of the original image according to its expanded bounding box $bbox_i' = (u_i', v_i', w_i', h_i')$:

$$M_i^{final}(x, y) = \begin{cases} M_i^{t^*}(x - u_i', y - v_i'), & (x, y) \in bbox_i', \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

where t^* denotes the final iteration index. The final forgery mask is obtained by merging all retained forged ROI masks:

$$M^{final} = \bigcup_{i \in \mathcal{S}_{forged}} M_i^{final}, \tag{14}$$

where $\mathcal{S}_{forged}$ denotes the index set of forged ROIs. In addition to the pixel-level IFL result, FGIR also outputs a structured IFL report that records the Confirmed ROIs, verified forensic clues, and refinement feedback.

Table 1: Comparison of IoU metrics. $\Delta\uparrow$ represents the performance improvement. The best results in each column are shown in **bold** and second best in underline.

Method	CASIAv1	Columbia	NIST16	Coverage	IMD2020	CocoGlide	SID-Set	Average
Training-based								
MVSS-Net++(Dong et al., 2022)	0.39	0.56	0.19	0.33	0.39	**0.40**	0.27	0.36
EVP(Liu et al., 2023b)	0.42	0.48	0.16	0.08	0.20	0.12	0.10	0.22
MPC(Lou et al., 2025)	0.41	**0.61**	**0.23**	0.30	**0.48**	0.32	0.28	0.38
Mesorch(Zhu et al., 2025)	0.49	0.59	0.22	**0.37**	0.40	0.36	**0.31**	**0.39**
PIM(Kong et al., 2025)	**0.51**	0.60	0.22	0.19	0.41	0.35	0.30	0.37
Zero-shot								
Llava-1.5-7B-hf(Liu et al., 2023a)	0.10	0.14	0.05	0.12	0.12	0.10	0.08	0.10
+ CLEAN	0.25	0.42	0.24	0.25	0.30	0.25	0.20	0.27
$\Delta\uparrow$	+0.15	+0.28	+0.19	+0.13	+0.18	+0.15	+0.12	+0.17
Llava-v1.6-Mistral-7B-hf(Liu et al., 2024)	0.12	0.14	0.08	0.09	0.14	0.12	0.10	0.11
+ CLEAN	0.29	0.47	0.33	0.27	0.35	0.30	0.25	0.32
$\Delta\uparrow$	+0.17	+0.33	+0.25	+0.18	+0.21	+0.18	+0.15	+0.21
GLM-4.1V-9B-Thinking(Hong et al., 2025)	0.14	0.15	0.07	0.08	0.15	0.13	0.10	0.12
+ CLEAN	0.34	0.49	0.28	0.29	0.38	0.32	0.28	0.34
$\Delta\uparrow$	+0.20	+0.34	+0.21	+0.21	+0.23	+0.19	+0.18	+0.22
Qwen2.5-VL-7B-Instruct(Team, 2025)	0.17	0.20	0.07	0.05	0.15	0.15	0.12	0.13
+ CLEAN	0.31	0.46	0.34	0.28	0.42	0.39	0.31	0.36
$\Delta\uparrow$	+0.14	+0.26	+0.27	+0.23	+0.27	+0.24	+0.19	+0.23
Qwen3-VL-8B-Instruct(Yang et al., 2025)	0.19	0.27	0.12	0.14	0.21	0.19	0.16	0.18
+ CLEAN	0.38	0.49	0.36	0.32	0.49	0.42	0.36	0.40
$\Delta\uparrow$	+0.19	+0.22	+0.24	+0.18	+0.28	+0.23	+0.20	+0.22
Qwen3.5-9B(Qwen Team, 2026)	0.23	0.33	0.16	0.24	0.26	0.24	0.19	0.23
+ CLEAN	**0.43**	**0.57**	**0.41**	**0.51**	**0.53**	**0.52**	**0.40**	**0.48**
$\Delta\uparrow$	+0.20	+0.24	+0.25	+0.27	+0.27	+0.28	+0.21	+0.25

Table 2: Comparison of F1 metrics. $\Delta\uparrow$ represents the performance improvement. The best results in each column are shown in **bold** and the second best in underline.

Method	CASIAv1	Columbia	NIST16	Coverage	IMD2020	CocoGlide	SID-Set	Average
Training-based								
MVSS-Net++(Dong et al., 2022)	0.51	0.66	0.23	**0.48**	0.38	**0.40**	**0.36**	0.43
EVP(Liu et al., 2023b)	0.48	0.61	0.21	0.11	**0.44**	0.15	0.12	0.30
MPC(Lou et al., 2025)	0.44	0.67	0.29	0.41	0.40	0.35	0.32	0.41
Mesorch(Zhu et al., 2025)	0.55	0.68	**0.39**	0.46	0.42	0.37	**0.36**	**0.46**
PIM(Kong et al., 2025)	**0.56**	**0.70**	0.35	0.39	0.34	0.38	0.33	0.43
Zero-shot								
Llava-1.5-7B-hf(Liu et al., 2023a)	0.25	0.24	0.16	0.21	0.14	0.12	0.10	0.17
+ CLEAN	0.37	0.51	0.45	0.37	0.36	0.30	0.25	0.37
$\Delta\uparrow$	+0.12	+0.27	+0.29	+0.16	+0.22	+0.18	+0.15	+0.20
Llava-v1.6-Mistral-7B-hf(Liu et al., 2024)	0.25	0.27	0.15	0.18	0.16	0.14	0.12	0.18
+ CLEAN	0.41	0.65	0.43	0.35	0.41	0.35	0.32	0.42
$\Delta\uparrow$	+0.16	+0.38	+0.28	+0.17	+0.25	+0.21	+0.20	+0.24
GLM-4.1V-9B-Thinking(Hong et al., 2025)	0.25	0.29	0.12	0.15	0.18	0.16	0.13	0.18
+ CLEAN	0.45	0.62	0.46	0.39	0.44	0.39	0.35	0.44
$\Delta\uparrow$	+0.20	+0.33	+0.34	+0.24	+0.26	+0.23	+0.22	+0.26
Qwen2.5-VL-7B-Instruct(Team, 2025)	0.29	0.31	0.12	0.09	0.26	0.26	0.21	0.22
+ CLEAN	0.42	0.65	0.47	0.37	0.51	0.54	0.42	0.48
$\Delta\uparrow$	+0.13	+0.34	+0.35	+0.28	+0.25	+0.28	+0.21	+0.26
Qwen3-VL-8B-Instruct(Yang et al., 2025)	0.32	0.39	0.20	0.25	0.24	0.23	0.19	0.26
+ CLEAN	0.46	0.68	0.48	0.42	0.55	0.56	0.46	0.52
$\Delta\uparrow$	+0.14	+0.29	+0.28	+0.17	+0.31	+0.33	+0.27	+0.26
Qwen3.5-9B(Qwen Team, 2026)	0.37	0.47	0.26	0.30	0.26	0.29	0.24	0.31
+ CLEAN	**0.53**	**0.74**	**0.59**	**0.55**	**0.61**	**0.63**	**0.49**	**0.59**
$\Delta\uparrow$	+0.16	+0.27	+0.33	+0.25	+0.35	+0.34	+0.25	+0.28

4 EXPERIMENTS

4.1 EXPERIMENTS SETUP

Datasets and Metrics. To comprehensively evaluate effectiveness and generalization, we conduct experiments on both traditional manipulation datasets, including CASIAv1 (Dong et al., 2013), Columbia (Ng et al., 2009), NIST16 (Guan et al., 2019), Coverage (Wen et al., 2016), IMD2020 (Novozamsky et al., 2020), and recent AIGC-based forgery benchmarks, including CocoGlide (Guillaro et al., 2023) and SID-Set (Huang et al., 2025). Consistent with previous works(Guillaro et al., 2023; Kong et al., 2025; Lou et al., 2025), the accuracy of pixel-level forgery

localization is evaluated using Intersection over Union (IoU) and F1-score. The configurations of the datasets and the details of the evaluation metrics are presented in Appendix C.

Implementation Details. To ensure a fair comparison, all methods are evaluated under a unified zero-shot evaluation protocol. For conventional training-based IFL methods, including MVSS-Net++ (Dong et al., 2022), EVP (Liu et al., 2023b), MPC (Lou et al., 2025), Mesorch (Zhu et al., 2025) and PIM (Kong et al., 2025), we directly use their officially released pretrained models. For MLLM-based methods, all MLLM and segmentation backbones are kept frozen. To minimize decoding stochasticity and improve reproducibility, we set the temperature to 0 for all MLLM inference. To comprehensively evaluate the effectiveness and generalization capability of CLEAN, we instantiate it with six different MLLM backbones. For each backbone, we further construct a zero-shot baseline, which directly applies the corresponding MLLM to the IFL task without the proposed CGSE and FGIR modules. This setting allows us to isolate the contribution of CLEAN under the same model prior. All key hyperparameters are kept consistent across experiments. We set the convergence threshold β to 0.99 and the maximum iteration number $T_{\max}$ to 3, while the remaining hyperparameter settings and sensitivity analyses are provided in Appendix D. All experiments are conducted on an NVIDIA RTX 4090 GPU with 24 GB memory.

4.2 Comparision results and analysis

The experimental evaluation was designed to validate the efficacy of the proposed CLEAN framework for the ZIFL task. The experimental design is robust, providing a comprehensive comparison between our CLEAN and both training-based SOTA methods and the MLLM baselines without the framework applied.

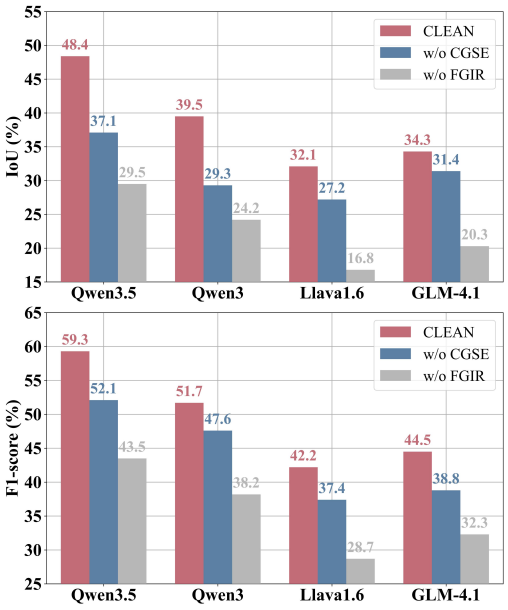


Figure 3: Ablation study on the contribution of CGSE and FGIR modules across different MLLMs.

As shown in Table 1 and Table 2,the CLEAN framework not only consistently and significantly outperforms the corresponding MLLM baselines across all test datasets , but also achieves average performance surpass training-based SOTA methods. More importantly, it demonstrates strong generalization across different manipulation patterns. Specifically, our CLEAN achieves substantial performance improvements not only on traditional manipulation datasets(*.e.g* Columbia (Ng et al., 2009),Coverage (Wen et al., 2016)), but also on AIGC-edit datasets (*.e.g* CocoGlide (Guillaro et al., 2023), SID-Set (Huang et al., 2025)). For instance, with Qwen3.5-9B as backbone, our CLEAN boosts the average IoU/F1 by **25%/28%** over its baseline, and its average performance of **48%/59%** surpasses SOTA methods like Mesorch (Zhu et al., 2025) (**39%/46%**) and PIM (Kong et al., 2025) (**37%/43%**). This can be attributed to the framework's effective synergy, which successfully gathers the cross-modal reasoning capacity of MLLMs and the fine-grained localization capacity of LangSAM.This performance is achieved by establishing a zero-shot collaborative framework that synergizes the MLLM's semantic understanding and LangSAM's precise segmentation capabilities, allowing it to identify forgery clues and iteratively refine results for pixel-level localization.

4.3 Ablation Studies

The ablation studies were conducted to quantify the distinct contributions of the framework's two core modules.

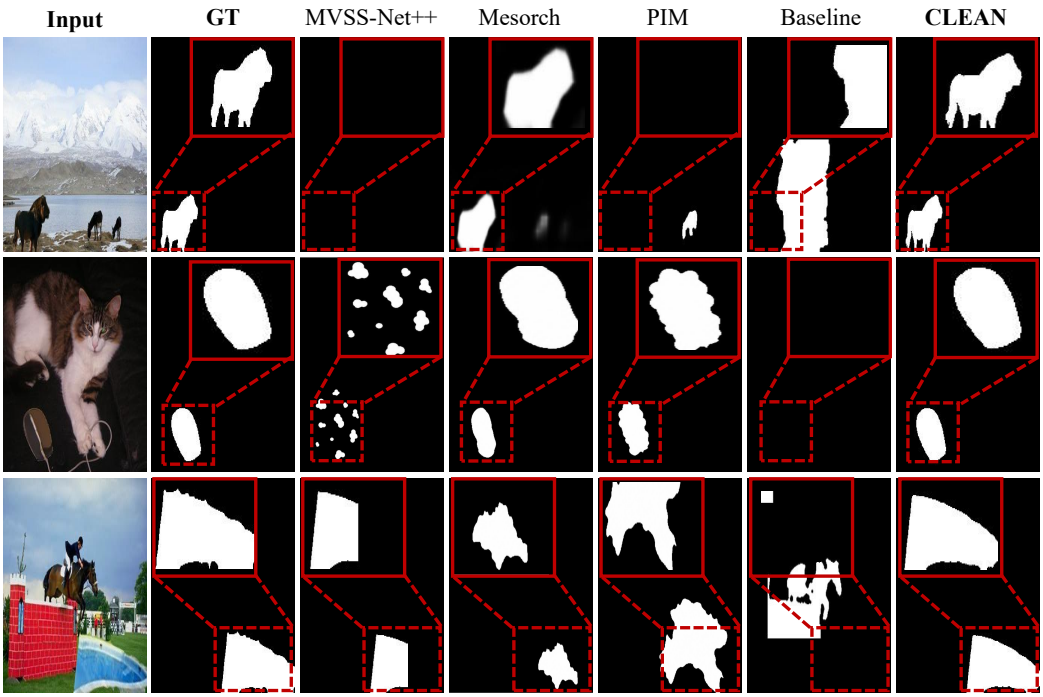


Figure 4: Comparisons between our CLEAN framework and other competitive methods.

Effectiveness of CGSE module. The results in Figure 3 consistently demonstrate that removing the CGSE module leads to a significant performance drop. Using Qwen3.5 as an example, the average IoU/F1 decreases from **48%/59%** to **37%/52%**. It is because that without this module, the framework fails to generate a high-recall list of candidate regions, which is crucial for the subsequent refinement stage. We further provide a fine-grained analysis of the three evidence dimensions in Appendix E.
Effectiveness of FGIR module. Removing the FGIR module results in a more severe performance degradation, causing a substantial drop of **19%** in average IoU and **16%** in average F1-score on the Qwen3.5. This can be attributed to the framework losing its critical ability to eliminate background noise and optimize the coarse boundaries provided by the initial stage. This loss underscores the module's essential function: iteratively judging forgery clues to progressively eliminate normal backgrounds and refine the boundaries of tampered regions.

4.4 Visualization

Figure 4 compares CLEAN with representative IFL methods. While existing methods often produce incomplete masks, fragmented predictions, or excessive background responses under different forgery manipulations, CLEAN consistently achieves more accurate and complete localization through multi-dimensional forensic reasoning and iterative refinement.

5 Conclusion

In this paper, We introduce CLEAN, a training-free framework for Zero-shot Image Forgery Localization (ZIFL). CLEAN addresses the limited generalization of training-based IFL methods under unseen forgery distributions by decomposing ZIFL into two coupled stages, combining Coarse-Grained Suspect Exploration (CGSE) for anomaly discovery with Fine-Grained Iterative Refinement (FGIR) for evidence verification and pixel-level boundary refinement. Experiments on traditional and AIGC-based forgery benchmarks demonstrate competitive localization performance and strong zero-shot generalization to unseen manipulation patterns.

Reproducibility Statement

The main paper describes the ZIFL setting, CLEAN framework, unified zero-shot evaluation protocol, inference settings, and key hyper-parameters. Appendix C provides the dataset configurations and evaluation metrics, while Appendix D reports the remaining hyper-parameter settings and sensitivity analyses. In addition, Appendix I provides the fixed prompt templates used by the CGSE and FGIR modules. These materials are intended to support reproduction of the reported results.

AI Use Statement

AI-assisted tools were used during manuscript preparation for language polishing and literature retrieval. They were not used to originate the research idea, formulate hypotheses, design the proposed framework or experimental protocol, analyze experimental results, or draw scientific conclusions. All technical content, experimental findings, claims, and conclusions were independently reviewed and verified by the authors, who take full responsibility for the final manuscript.

References

Jawadul H Bappy, Cody Simons, Lakshmanan Nataraj, BS Manjunath, and Amit K Roy-Chowdhury. Hybrid lstm and encoder–decoder architecture for detection of image forgeries. *IEEE transactions on image processing*, 28(7):3286–3300, 2019.

Xiuli Bi, Yang Wei, Bin Xiao, and Weisheng Li. Rru-net: The ringed residual u-net for image splicing forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Tiago Carvalho, Fabio A Faria, Helio Pedrini, Ricardo da S Torres, and Anderson Rocha. Illuminant-based transformed spaces for image forensics. *IEEE transactions on information forensics and security*, 11(4): 720–733, 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, 2022.

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. IEEE, 2013.

Pasquale Ferrara, Tiziano Bianchi, Alessia De Rosa, and Alessandro Piva. Image forgery localization via fine-grained analysis of cfa artifacts. *IEEE Transactions on Information Forensics and Security*, 7(5):1566–1577, 2012.

Zan Gao, Chao Sun, Zhiyong Cheng, Weili Guan, Anan Liu, and Meng Wang. Tbnet: A two-stream boundary-aware network for generic image manipulation localization. *IEEE Transactions on knowledge and data engineering*, 35(7):7541–7556, 2022.

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 *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 63–72. IEEE, 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.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *European conference on computer vision*, pp. 312–328. Springer, 2020.

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 Computer Vision and Pattern Recognition Conference*, pp. 28831–28841, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36:29914–29934, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Chenqi Kong, Anwei Luo, Shiqi Wang, Haoliang Li, Anderson Rocha, and Alex C Kot. Pixel-inconsistency modeling for image manipulation localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Jaeseong Lee, Yeeun Choi, Heechan Choi, Hanjung Kim, and Seonjoo Kim. A training-free, task-agnostic framework for enhancing mllm performance on high-resolution images. *arXiv preprint arXiv:2507.10202*, 2025.

Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9098–9108, 2025.

Haodong Li and Jiwu Huang. Localization of deep inpainting using high-pass fully convolutional network. In *proceedings of the IEEE/CVF international conference on computer vision*, pp. 8301–8310, 2019.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19434–19445, 2023b.

Zijie Lou, Gang Cao, Kun Guo, Lifang Yu, and Shaowei Weng. Exploring multi-view pixel contrast for general and robust image forgery localization. *IEEE Transactions on Information Forensics and Security*, 20: 2329–2341, 2025.

Babak Mahdian and Stanislav Saic. Using noise inconsistencies for blind image forensics. *Image and vision computing*, 27(10):1497–1503, 2009.

Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, and Federico Tombari. I2mvformer: Large language model generated multi-view document supervision for zero-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15169–15179, 2023.

TT Ng, SF Chang, and Q Sun. Columbia uncompressed image splicing detection evaluation dataset. *Columbia Univ. ADVENT Tech. Rep.*, 2009.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

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.

Jeong-Woo Park and Seong-Whan Lee. Mcot-re: Multi-faceted chain-of-thought and re-ranking for training-free zero-shot composed image retrieval. *arXiv preprint arXiv:2507.12819*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Yuan Rao, Jiangqun Ni, Weizhe Zhang, and Jiwu Huang. Towards jpeg-resistant image forgery detection and localization via self-supervised domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 47(5):3285–3297, 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ke Sun, Shen Chen, Taiping Yao, Ziyin Zhou, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. Towards general visual-linguistic face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19576–19586, 2025.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Wenbin Wang, Yongcheng Jing, Liang Ding, Yingjie Wang, Li Shen, Yong Luo, Bo Du, and Dacheng Tao. Retrieval-augmented perception: High-resolution image perception meets visual rag. In *International Conference on Machine Learning*, pp. 63290–63307. PMLR, 2025.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36:61501–61513, 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 *2016 IEEE international conference on image processing (ICIP)*, pp. 161–165. Ieee, 2016.

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. *arXiv preprint arXiv:2410.02761*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

S Zhang, P Sun, S Chen, M Xiao, W Shao, W Zhang, Y Liu, K Chen, and P Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. arxiv 2023. *arXiv preprint arXiv:2307.03601*, 2023.

Yue Zhang, Ben Colman, Xiao Guo, Ali Shahriyari, and Gaurav Bharaj. Common sense reasoning for deepfake detection. In *European conference on computer vision*, pp. 399–415. Springer, 2024.

Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1053–1061, 2018.

Peng Zhou, Bor-Chun Chen, Xintong Han, Mahyar Najibi, Abhinav Shrivastava, Ser-Nam Lim, and Larry Davis. Generate, segment, and refine: Towards generic manipulation segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13058–13065, 2020.

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 & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 11022–11030, 2025.

Peiyu Zhuang, Haodong Li, Shunquan Tan, Bin Li, and Jiwu Huang. Image tampering localization using a dense fully convolutional network. *IEEE Transactions on Information Forensics and Security*, 16:2986–2999, 2021.

A THEORETICAL JUSTIFICATION

A.1 CONTEXT-AWARE REGION EXPANSION IN CGSE MODULE.

In Section 3.3, we introduce the dynamic region expansion operation to enlarge the coarse bounding box generated by the MLLM. This operation is designed to compensate for the incomplete spatial coverage of MLLM-generated proposals, since the MLLM tends to locate the central suspicious object while ignoring boundary-level forged traces.

Given the i-th suspicious proposal, the original bounding box is denoted as $bbox_i = (x_i, y_i, w_i, h_i)$. Let G_i denote the corresponding real forged region. Since the MLLM generates $bbox_i$ mainly according to high-level semantic clues, we assume that the maximum spatial deviation between $bbox_i$ and G_i is bounded by

$$\Delta_i \leq \alpha_i \cdot \max(w_i, h_i), \tag{15}$$

According to the dynamic region expansion operation, the expanded bounding box is computed as

$$bbox'_i = (x_i - \alpha_i w_i,\ y_i - \alpha_i h_i,\ (1 + 2\alpha_i)w_i,\ (1 + 2\alpha_i)h_i). \tag{16}$$

When the expansion ratio α_i satisfies the above deviation bound, the potential forged region can be covered by the expanded proposal:

$$G_i \subseteq bbox'_i. \tag{17}$$

Therefore, the region-level recall of the expanded proposal satisfies

$$\text{Recall}(bbox'_i, G_i) = \frac{|bbox'_i \cap G_i|}{|G_i|} = 1. \tag{18}$$

In contrast, the original coarse bounding box may miss boundary-level forged traces, leading to

$$\text{Recall}(bbox_i, G_i) = \frac{|bbox_i \cap G_i|}{|G_i|} < 1. \tag{19}$$

This derivation shows that the context-aware expansion mechanism improves the coverage of potential forged regions. Therefore, it provides a high-recall proposal list for the following fine-grained refinement stage.

A.2 CRITIC-GUIDED BOUNDARY EVOLUTION

In Section 3.4, we introduce the Critic-Guided Boundary Evolution mechanism to refine the ROI mask by iteratively updating the segmentation prompt according to forensic consistency. This process aims to progressively correct inaccurate boundaries, remove irrelevant background regions, and align the predicted mask with the verified forensic evidence.

Let M_i^* denote the ideal forgery mask inside the confirmed ROI R_i, and let M_i^t denote the mask generated at the t-th refinement round. The mask is produced by LangSAM with the current prompt Q_i^t:

$$M_i^t = \text{LangSAM}(R_i, Q_i^t). \tag{20}$$

We define the mask error as

$$e_t = 1 - \text{IoU}(M_i^t, M_i^*). \tag{21}$$

At each refinement round, the Critic evaluates the current mask with respect to the confirmed ROI and the verified forensic evidence, and provides feedback a_i^t for prompt refinement:

$$a_i^t = \text{Critic}(R_i, M_i^t, \mathcal{E}_i), \tag{22}$$

where $\mathcal{E}_i = \{e_j \mid v_j = \text{Support}\}$ denotes the set of supported evidence descriptions. The next prompt is then updated as

$$Q_i^{t+1} = \text{Refine}(Q_i^t, a_i^t). \tag{23}$$

To analyze the effect of this iterative refinement, we assume that each valid critic-guided update reduces the previous mask error by a factor ρ, where $0 < \rho < 1$, up to a small residual error ϵ caused by ambiguous visual evidence:

$$e_{t+1} \leq \rho e_t + \epsilon. \tag{24}$$

By recursively applying the above inequality, we obtain

$$e_t \leq \rho^t e_0 + \epsilon \sum_{j=0}^{t-1} \rho^j = \rho^t e_0 + \frac{1-\rho^t}{1-\rho}\epsilon. \tag{25}$$

Since $0 < \rho < 1$, the term $\rho^t e_0$ decreases as the iteration number increases. Thus, the mask error is bounded by

$$\lim_{t\to\infty} e_t \leq \frac{\epsilon}{1-\rho}. \tag{26}$$

This indicates that Critic-Guided Boundary Evolution can progressively reduce the mask error until the refinement reaches a stable state, with the residual error determined by the ambiguity of the visual evidence.

Moreover, CLEAN adopts an IoU-based convergence criterion:

$$\mathrm{IoU}(M_i^t, M_i^{t-1}) > \beta. \tag{27}$$

When this condition is satisfied, two adjacent masks become sufficiently similar, indicating that further prompt updates bring limited changes to the segmentation result. Therefore, the iteration can be stopped without introducing unnecessary refinement rounds.

A.3 Final-Judgment and Localization Results Generation

In Section 3.4, after boundary evolution, each confirmed forged ROI produces a refined local mask. The final stage maps these local masks back to the original image coordinate system and integrates them into an image-level forgery localization result.

Let $\mathcal{S}_{forged}$ denote the index set of ROIs finally judged as forged after FGIR. For each $i \in \mathcal{S}_{forged}$, the refined local mask is mapped back to the original image coordinate system as M_i^{final}. The final image-level forgery mask is obtained by the union of all confirmed forged ROI masks:

$$M^{final} = \bigcup_{i\in\mathcal{S}_{forged}} M_i^{final}. \tag{28}$$

This union-based integration ensures that multiple confirmed forged ROIs can be preserved simultaneously in the final localization result.

Let the complete ground-truth forgery mask be composed of multiple forged components:

$$G = \bigcup_{k=1}^{K} G_k. \tag{29}$$

Assume that each forged component G_k is covered by at least one expanded proposal after context-aware region expansion, confirmed after final judgment, and refined into a valid local mask. Formally, for each G_k, there exists an index $i_k \in \mathcal{S}_{forged}$ such that

$$G_k \cap M_{i_k}^{final} \neq \emptyset. \tag{30}$$

Then, since $M_{i_k}^{final}$ is included in the final union, we have

$$M_{i_k}^{final} \subseteq M^{final}. \tag{31}$$

Therefore, all confirmed positive local masks are preserved in M^{final}:

$$\bigcup_{i\in\mathcal{S}_{forged}} M_i^{final} = M^{final}. \tag{32}$$

This derivation shows that the final integration stage preserves multiple detected forged regions by merging all positive local masks. Meanwhile, ROIs classified as authentic or unsupported are filtered out before integration, which reduces false-positive background regions introduced by the high-recall exploration stage.

Overall, the above derivations justify the key operations of CLEAN. The context-aware region expansion improves the coverage of potential forged regions, the Critic-Guided Boundary Evolution progressively reduces mask errors through forensic-consistency feedback, and the Final-Judgment and Localization Results Generation stage preserves all confirmed forged ROI masks while filtering out unreliable candidates.

Algorithm 1: Fine-Grained Iterative Refinement

Input: Image x_i, expanded proposal set $\mathcal{P}' = \{p_i'\}_{i=1}^{N_p}$
Output: Final forgery mask M^{final} and structured IFL report $\mathcal{R}^{final}$
$M^{final} \leftarrow \varnothing, \mathcal{R}^{final} \leftarrow \varnothing, \mathcal{S}_{confirmed} \leftarrow \varnothing$;
`// Confidence-aware Pre-Judgment`
foreach $p_i' \in \mathcal{P}'$ **do**
  Crop the candidate ROI R_i from x_i according to $bbox_i'$;
  $u_i \leftarrow \text{MLLM_Verify}(R_i,\, clues_i)$;
  if $u_i ==$ ***False*** **then**
    Label R_i as "Authentic";
    continue;
  $score_i \leftarrow \lambda \cdot S_{\text{prior}}(conf_i) + (1-\lambda) \cdot \frac{|clues_i|}{\max_j |clues_j|}$;
  Add $(R_i, p_i', score_i)$ to $\mathcal{S}_{confirmed}$;
Sort $\mathcal{S}_{confirmed}$ by $score_i$ in descending order;
foreach $(R_i, p_i', score_i) \in \mathcal{S}_{confirmed}$ **do**
  `// Hierarchical Evidence Decomposition`
  $\{T_j\}_{j=1}^{K_i} \leftarrow \text{MLLM}_{\text{Leader}}(R_i,\, clues_i)$;
  $\mathcal{Q}_i \leftarrow \varnothing, \mathcal{E}_i \leftarrow \varnothing$;
  foreach T_j **do**
    $(v_j, e_j, q_j) \leftarrow \text{MLLM}_{\text{Worker}_j}(R_i,\, T_j)$;
    if $v_j ==$ Support **then**
      $\mathcal{Q}_i \leftarrow \mathcal{Q}_i \cup \{q_j\}$;
      $\mathcal{E}_i \leftarrow \mathcal{E}_i \cup \{e_j\}$;
  if $\mathcal{Q}_i == \varnothing$ **then**
    Label R_i as "Authentic";
    continue;
  $Q_i^0 \leftarrow \text{Aggregate}(\mathcal{Q}_i)$;
  `// Critic-Guided Boundary Evolution`
  $t \leftarrow 0, M_i^{-1} \leftarrow \varnothing$;
  while $t < T_{\max}$ **do**
    $M_i^t \leftarrow \text{LangSAM}(R_i,\, Q_i^t)$;
    $a_i^t \leftarrow \text{Critic}(R_i,\, M_i^t,\, \mathcal{E}_i)$;
    if $a_i^t ==$ *"Good Enough"* **then**
      break;
    if $t > 0$ ***and*** $\text{IoU}(M_i^t, M_i^{t-1}) > \beta$ **then**
      break;
    $Q_i^{t+1} \leftarrow \text{Refine}(Q_i^t,\, a_i^t)$;
    $t \leftarrow t+1$;
  `// Final-Judgment and Localization Results Generation`
  if $M_i^t \neq \varnothing$ **then**
    Map M_i^t back to the original image coordinate system according to $bbox_i'$ to obtain M_i^{final};
    $M^{final} \leftarrow M^{final} \cup M_i^{final}$;
    $\mathcal{R}^{final} \leftarrow \mathcal{R}^{final} \cup \{R_i, \mathcal{E}_i, a_i^{0:t}\}$;

B ALGORITHM OF PROPOSED APPROACH

The detailed procedure of the proposed Fine-Grained Iterative Refinement (FGIR) module is shown in Algorithm 1.

C Datasets and Evaluation Metrics.

C.1 Datasets

To evaluate the proposed CLEAN framework under the Zero-shot Image Forgery Localization (ZIFL) setting, we conduct experiments on seven representative image forgery localization benchmarks, including CASIAv1 (Dong et al., 2013), Columbia (Ng et al., 2009), NIST16 (Guan et al., 2019), Coverage (Wen et al., 2016), IMD2020 (Novozamsky et al., 2020), CocoGlide (Guillaro et al., 2023), and SID-Set (Huang et al., 2025). These datasets cover diverse manipulation paradigms, ranging from traditional copy-move and splicing operations to recent AIGC-based image editing forgeries. Such a benchmark design allows us to comprehensively verify whether CLEAN can generalize not only to conventional manipulation traces, but also to emerging AI-generated editing artifacts without any additional training. Specifically, these benchmarks can be divided into two categories:

Traditional manipulation datasets. The traditional manipulation group consists of CASIAv1, Columbia, NIST16, Coverage, and IMD2020, covering representative conventional forgeries such as copy-move, splicing, region insertion, and post-processed editing. CASIAv1 and Columbia mainly focus on object-level splicing and region insertion, while Coverage emphasizes copy-move forgeries. NIST16 and IMD2020 contain more diverse and realistic manipulations with complex post-processing. These datasets evaluate CLEAN on traditional forgery traces, including boundary artifacts, texture discontinuity, duplicated structures, and cross-region inconsistency.

AIGC-editing datasets. The AIGC-editing group consists of CocoGlide and SID-Set, which evaluate CLEAN on recent generative editing forgeries. Compared with traditional manipulations, AIGC-edited regions usually have more natural boundaries and coherent local textures, making localization more challenging. CocoGlide is generated from COCO images using the GLIDE (Nichol et al., 2021) diffusion model through generative inpainting. SID-Set contains social-media-oriented AIGC forgeries generated with advanced generative and editing models, including GPT-4o (Hurst et al., 2024), Language-SAM, Latent Diffusion (Rombach et al., 2022), and FLUX (Labs et al., 2025).

This evaluation protocol is consistent with the core motivation of our CLEAN: a practical zero-shot forgery localization method should be able to localize manipulated regions across both historical forgery types and newly emerging generative editing techniques without relying on task-specific training data.

C.2 Evaluation Metrics

Following previous image forgery localization works, we evaluate the pixel-level localization performance using Intersection over Union (IoU) and F1-score. Given the predicted forgery mask $\hat{Y}$ and the ground-truth mask Y, IoU is defined as:

$$IoU = \frac{|\hat{Y} \cap Y|}{|\hat{Y} \cup Y|}. \tag{33}$$

Equivalently, IoU can be computed based on true positives (TP), false positives (FP), and false negatives (FN):

$$IoU = \frac{TP}{TP + FP + FN}. \tag{34}$$

The F1-score is the harmonic mean of precision and recall:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \tag{35}$$

where

$$Precision = \frac{TP}{TP + FP}, \quad Recall = \frac{TP}{TP + FN}. \tag{36}$$

Thus, the F1-score can also be written as:

$$F1 = \frac{2TP}{2TP + FP + FN}. \tag{37}$$

D HYPER-PARAMETER ANALYSIS

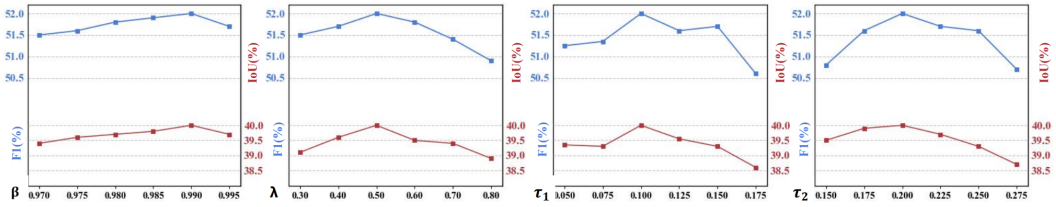


Figure 5: The influence of hyper-parameters in our CLEAN, using Qwen3-VL-8B-Instruct as the backbone.

Our proposed CLEAN involves four key hyper-parameters: β, λ, τ_1, and τ_2. We individually analyze their effects as shown in Fig. 5. Specifically, CLEAN achieves the best performance when β and λ are set to 0.990 and 0.50, respectively, indicating that a strict convergence criterion and a balanced proposal scoring strategy are beneficial for stable forgery localization. In addition, as τ_1 and τ_2 increase, the performance first improves and then decreases, reaching the optimum at 0.100 and 0.200, respectively.This is because moderate region-expansion thresholds can preserve sufficient suspicious context while avoiding excessive normal backgrounds. Therefore, we set β, λ, τ_1, and τ_2 to 0.990, 0.50, 0.100, and 0.200 in all experiments.

Table 3: Ablation study on the contributions of the multi-dimensional evidence search (Semantic Consistency, Digital Integrity, Physical Environment Consistency) across comprehensive datasets, using Qwen3-VL-8B-Instruct as the backbone.

Method	CASIAv1		Columbia		NIST16		Coverage		IMD2020		CocoGlide		SID-Set		Average	
	Iou	F1	Iou	F1	Iou	F1	Iou	F1	Iou	F1	Iou	F1	Iou	F1	Iou	F1
Qwen3-VL	0.19	0.32	0.27	0.39	0.12	0.20	0.14	0.25	0.21	0.24	0.19	0.23	0.16	0.19	0.18	0.26
+ Semantic	0.27	0.38	0.38	0.51	0.18	0.28	0.18	0.28	0.31	0.35	0.28	0.34	0.24	0.29	0.26	0.35
+ Digital	0.31	0.42	0.41	0.54	0.25	0.34	0.26	0.34	0.35	0.41	0.30	0.38	0.26	0.32	<u>0.31</u>	<u>0.39</u>
+ Phy-Env.	0.26	0.34	0.34	0.48	0.25	0.36	0.22	0.31	0.30	0.36	0.29	0.35	0.25	0.30	0.27	0.36
+ CLEAN	0.38	0.46	0.49	0.68	0.36	0.48	0.32	0.42	0.49	0.55	0.42	0.56	0.36	0.46	**0.40**	**0.52**

E ABLATION ON EVIDENCE DIMENSIONS

In this section, we analyzed the additive contribution of the three evidence dimensions within the CGSE module (Table 3). While each dimension provides a clear individual boost over the baseline, synergizing all three dimensions simultaneously yields the most significant performance improvement, which boosts the average IoU by **22%** and the F1-score by **26%**. This result underscores the module's critical role in exploiting the reasoning potential of large models to discover suspected area based on multi-dimensional clues.

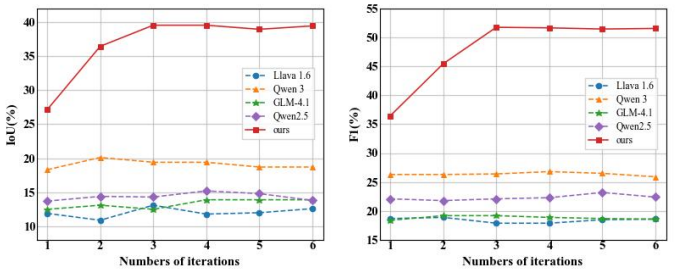


Figure 6: Impact of the number of iterations (T_{max}) on modelperformance

F Efficiency and Comparative Analysis

A major concern for MLLM-based forgery localization is that large-model reasoning may introduce additional inference cost. We emphasize that CLEAN is designed for a zero-shot forgery localization setting, where no task-specific training data or model updating is required. As shown in Table 4, we further evaluate the efficiency of CLEAN under the same MLLM backbone. Taking Qwen3.5 as the backbone, the corresponding baseline requires 12.9s per image and obtains **23%/31%** (IoU/F1). By introducing CLEAN, the inference time increases to 18.3s per image, while the performance is improved to **48%/59%** (IoU/F1). This comparison indicates that the additional inference cost mainly comes from the zero-shot collaborative reasoning and iterative refinement process, and is effectively converted into substantial localization gains. Figure 6 further analyzes the influence of the maximum iteration number $T_{\max}$. When the number of iterations increases from 1 to 3, the performance of CLEAN improves by **13%/16%**, demonstrating that the iterative refinement process can progressively correct coarse masks and remove irrelevant background regions.

These observations also explain why CLEAN adopts an iterative zero-shot reasoning paradigm. Compared with a naive MLLM baseline, CLEAN introduces several additional reasoning steps, but these steps are targeted at proposal verification, prompt refinement, and boundary correction. As a result, CLEAN achieves a better trade-off between inference cost and localization quality. Therefore, although CLEAN is slightly slower than a single-pass MLLM baseline during inference, it provides higher overall efficiency in open-set and rapidly evolving forgery localization scenarios, where zero-shot generalization is more critical than single-image latency alone.

Table 4: Efficiency comparison of CLEAN with other MLLMs.

Method	Infer time / img	IoU	F1
Llava 1.6	**9.6** s	0.11	0.18
GLM-4.1	11.4 s	0.12	0.18
Qwen3	14.3 s	0.18	0.26
Qwen3.5	12.9s	0.23	0.31
Qwen3.5+CLEAN	18.3 s	**0.48**	**0.59**

G Robustness to Signal-level Artifacts

Though MLLMs are natively designed for semantic awareness, they often struggle with low-level spectral variations. We bridge this gap through customized prompts in our Coarse-Grained Suspect Exploration (CGSE) module. By explicitly prompting for *"noise patterns"* and *"abnormal edges"* in digital integrity, we guide CLEAN to focus on signal-level artifacts. Unlike training-based CNN method, our CLEAN transfers the semantic perceive capacity of MLLMs to spectral signal statistics via iterative forgery localization with customized prompts in a training-free way. As demonstrated in Table 5, removing prompts for noise and abnormal edges significantly degrades performance. Our full CLEAN pipeline improves by **8%/9%** w/ *"noise"*/*"ab edges"*, proving its robustness to spectrally forged signals.

Table 5: Ablation study on signal-level customized prompts using Qwen3-VL-8B-Instruct as the backbone. Results are reported in terms of F1-score.

Dataset	w/o noise & ab edges	w/o ab edges	w/o noise	CLEAN
Columbia (Ng et al., 2009)	0.51	0.55	0.62	**0.68**
CocoGlide (Guillaro et al., 2023)	0.41	0.51	0.46	**0.56**
Average	0.46	0.53	0.54	**0.62**

H Limitation Discussion

Although CLEAN achieves promising zero-shot localization performance without task-specific parameter updating, it still has several limitations. First, the iterative refinement process introduces additional inference cost compared with single-pass localization methods. Second, the quality of the final mask is affected by the accuracy of the generated forensic clues and segmentation prompts, especially when manipulated regions are extremely small, visually ambiguous, or heavily blended with the surrounding context. We will explore more efficient reasoning strategies in future work.

I Implementation Details of Prompt-Guided Forensic Analysis

This section provides the prompt-level implementation details of the CGSE module and the FGIR module in CLEAN. The Forensic Analyzer in CGSE and the three specialist (Leader, Worker and Critic) in FGIR are all implemented through fixed prompt templates Specifically, we detail the prompts used for Coarse-Grained Suspect Exploration (CGSE), Leader-based task decomposition, Worker-based sub-region verification, Chief-based prompt aggregation, and Critic-based iterative refinement.

Coarse-Grained Suspect Exploration. As shown in Figure 2, CGSE starts from the input image and establishes a scene-level baseline from the image caption. Then we prompt the MLLM with the identity of a Forensic Analyzer to search for suspicious ROIs from multiple forensic dimensions, including semantic consistency, digital integrity, and physical-environment consistency. The output is a structured candidate list rather than a single image-level answer. Each candidate ROI contains a description, a bounding box, a suspicion level, and a set of forensic clues. Listing 1 summarizes the prompt template used for this stage.

Listing 1 Prompt for Forensic Analyzer.

```
"""You are a professional image forgery analyst.

Input:
- The input image.
- A global scene caption: {caption}.
- Image size: {image_dimensions}.

Your task:
Scan the whole image and report all potentially suspicious regions.
    Prefer high
recall: do not miss plausible forged regions, even if some candidates are
     later
rejected by the fine-grained verifier.

Analysis protocol:
1. Establish a scene baseline from the caption and the visible image.
2. Search for suspicious regions from the following dimensions:
   - Semantic Consistency: objects, relations, or events that violate the
     scene context or common sense.
   - Digital Integrity: unnatural boundaries, jagged edges, texture
    discontinuities, inconsistent noise, or compression mismatch.
   - Physical-Environment Consistency: inconsistent lighting, missing
    shadows, wrong perspective, abnormal scale, or implausible contact.
3. Group related clues into candidate ROIs. Each ROI should correspond to
     one core physical entity whenever possible.
4. Estimate a tight bounding box for each ROI.

Output strictly as a JSON object:
{
  "1": {
    "description": "A short description of the suspicious object/region.",

    "bounding_box": [x_min, y_min, x_max, y_max],
    "suspicion_level": "High" or "Medium" or "Low",
    "clues": [
      "Semantic clue ...",
      "Digital integrity clue ...",
      "Physical-environment clue ..."
    ]
  }
}
"""
```

Leader-based task decomposition. After CGSE produces an expanded candidate ROI, the Leader receives the cropped ROI and its clues. Its role is to convert the high-level suspicion into a small set of executable verification tasks. In our implementation, the Leader selects at most three relevant forensic dimensions from boundary integrity, noise/texture consistency, lighting plausibility, and semantic logic. Listing 2 gives the corresponding template.

Worker-based sub-region verification. Each Worker executes one verification task on the cropped ROI. The Worker is not allowed to judge the whole image or introduce new unrelated hypotheses. It only checks whether the assigned clue is visually supported inside the sub-region. If the clue is supported, the Worker must also generate a LangSAM-compatible segmentation prompt that describes the concrete object to segment, rather than an abstract defect. This design directly connects forensic reasoning to pixel-level localization. Listing 3 shows the Worker prompt template.

Listing 2 Prompt for Leader-based verification task decomposition.

```
"""You are a chief forensic task planner.

Input:
- A cropped suspicious ROI from the original image.
- Candidate forensic clues: {clues}.

Your task:
Convert the coarse suspicious clues into a small set of concrete,
    object-
centered verification tasks for Workers.

Available forensic dimensions:
- Boundary Integrity: check hard cut edges, jagged boundaries, halo
    artifacts, abnormal blur, or stitching traces.
- Noise/Texture Consistency: check noise pattern mismatch, repeated
    texture, cloning traces, or content-fill artifacts.
- Lighting Plausibility: check light direction, shadow existence,
    shadow shape, highlight, reflection, and color temperature.
- Semantic Logic: check perspective, scale, physical interaction,
    object relation, and scene common sense.

Task generation rule:
Select at most three dimensions that are most relevant to the clues.
    For each
task, identify the suspicious physical object and specify what the
    Worker
should verify.

Output strictly as JSON:
{
  "tasks": [
    {
      "task_id": "task_001",
      "verification_dimension": "Boundary Integrity",
      "task_content": "Inspect the boundary of the suspicious object
    and determine whether it shows unnatural stitching or cut
    artifacts."
    }
  ]
}
"""
```

Chief-based evidence aggregation and prompt selection. The Workers may produce multiple candidate prompts because different forensic dimensions can point to related but not identical visual entities. Therefore, CLEAN uses a Chief aggregation prompt to cross-check the Worker reports. If all Workers return *Reject*, the ROI is removed. Otherwise, the Chief selects the final LangSAM

prompt from the Worker-proposed candidates according to evidence consistency and subject-context separation. Listing 4 shows this aggregation template.

LangSAM-driven mask generation and Critic-based refinement. The selected prompt is passed to LangSAM to obtain an initial mask. When multiple prompts remain, CLEAN segments them separately and merges the masks. For later iterations, the Critic evaluates whether the current mask is accurate enough. The Critic does not re-assess authenticity; it only checks whether the frozen forensic evidence is fully covered by the mask. If the mask is incomplete or polluted by unrelated background, the Critic generates an optimized object-level prompt for another LangSAM pass. Listing 5 gives the Critic template.

Listing 3 Prompt for Worker-based sub-region forensic analysis.

```
"""You are a forensic image analysis Worker.

Input:
- A cropped suspicious sub-region from the original image.
- The original suspicious clues: {clues}.
- An assigned forensic clue: {clue}.
- An assigned forensic dimension: {dimension}.
- A concrete verification task: {task_content}.

Your task:
Analyze only the visible content inside the cropped sub-region. Do
    not judge
the whole image. Do not assume the region is forged unless the visual
     evidence
is clear.

Inspection criteria:
- If the dimension is Semantic Consistency, check whether the object,
     object relation, or scene context is semantically implausible.
- If the dimension is Digital Integrity, check whether there are
    unnatural boundaries, texture discontinuities, inconsistent noise,
     compression mismatch, or pixel-level artifacts.
- If the dimension is Physical-Environment Consistency, check whether
     lighting, shadow, reflection, perspective, scale, or spatial
    layout is  physically inconsistent.

Decision rule:
Return "Support" only when the assigned clue is clearly visible and
    cannot be
reasonably explained by normal imaging or natural scene variations.
Return "Reject" when the clue is weak, ambiguous, invisible, or
    explainable
by normal visual conditions.

Segmentation prompt rule: If the verdict is "Support", generate a
    short English
phrase that describes the concrete visible object to be segmented by
    LangSAM.
Describe the object itself, not the defect. For example, use "the
    yellow duck"
instead of "the jagged edge"; use "the wallet" instead of "the
    abnormal bottom
boundary". If multiple same-category objects exist, use a simple
    position or
color modifier, e.g., "the red car on the left". If the verdict is "
    Reject",
set the segmentation prompt to an empty string.

Output strictly in the following JSON format:
{
  "verdict": "Support" or "Reject",
  "evidence": "A concise description of the concrete visual evidence
    .",
  "segmentation_prompt": "A short phrase that can guide LangSAM to
    segment the suspicious area."
}

Format: <verdict> <evidence> <segmentation_prompt> STOP
"""
```

Listing 4 Prompt for Chief-based Worker aggregation.

```
"""You are a chief forensic examiner.

Input:
- The cropped suspicious ROI.
- Positive Worker findings: {positive_findings}.
- Candidate LangSAM prompts proposed by Workers: {candidate_prompts}.

Your task:
Aggregate the Worker reports and select the final segmentation target.


Reasoning principles:
1. Evidence consistency: if multiple forensic dimensions point to the
     same physical object, that object is the most likely forged
    subject.
2. Subject-context separation: distinguish the forged subject from
    the contextual evidence that reveals the forgery. For example, an
    abnormal shadow may reveal the forgery, but the segmentation
    target should be the object casting the shadow.

Output rule:
Select the prompt or prompts that best represent the core forged
    subject. The
selected prompts must come from the candidate prompt list.

Output strictly as JSON:
{
  "overall_conclusion": "A concise explanation of the selected forged
     subject.",
  "definitive_sam_prompts": ["the selected prompt from
    candidate_prompts"]
}
"""
```

J BROADER IMPACTS

Our work studies Zero-shot Image Forgery Localization (ZIFL) and proposes CLEAN to localize manipulated regions without task-specific parameter updating. The potential positive societal impacts of this research include improving digital media trustworthiness and supporting practical forensic analysis under rapidly evolving manipulation techniques. As image editing tools and generative models become increasingly accessible, manipulated images are more diverse and realistic. CLEAN can help identify and localize suspicious regions at the pixel level, which may benefit misinformation verification, media forensics, copyright protection, and trustworthy visual content analysis. In addition, by avoiding task-specific training on each target forgery distribution, CLEAN reduces the dependence on newly collected annotated forgery data and provides a more flexible solution for unseen manipulation patterns.

This work may also introduce potential risks. First, stronger forgery localization systems may provide malicious users with feedback about which manipulated regions are easier to detect, potentially helping them improve evasion strategies. Second, like other automated forensic tools, CLEAN may still produce false positives or false negatives, especially for extremely subtle, low-quality, or heavily post-processed images. Therefore, the predictions of CLEAN should be used as supportive evidence rather than the sole basis for high-stakes decisions in journalism, law enforcement, or legal investigations.

Overall, CLEAN is expected to contribute positively to trustworthy media analysis and robust image forensics. Meanwhile, careful deployment, human verification, and continuous evaluation are necessary to mitigate the risks of misuse and over-reliance on automated forensic predictions.

Listing 5 Prompt for Critic-guided boundary evaluation.

```
"""You are a chief digital forensic analyst specialized in precise
forgery localization.

Input:
- The current masked sub-region.
- The original LangSAM prompts: {original_prompts}.
- Frozen forensic clues from the discovery stage: {clues}.

Your task:
Do not re-judge whether the region is forged. Assume the frozen clues
    are valid
and evaluate only whether the current mask accurately covers the
    forensic subject.

Evaluation criteria:
1. Target completeness: does the mask fully cover the original target?

2. Evidence-chain completeness: does the mask include all visible
    traces related to the frozen clues?
3. Boundary closure: does the mask follow the outer contour of the
    forensic subject without obvious missing parts?
4. Range pollution: does the mask include excessive unrelated normal
    background or independent irrelevant objects?

Decision rule:
Return "Good Enough" if the target and evidence-related traces are
    covered with
no major omission or severe pollution. Return "Needs Improvement" if
    the target
is incomplete, important evidence is missing, boundaries are
    inaccurate, or
unrelated background is heavily included.

Optimized prompt rule:
Only when the conclusion is "Needs Improvement", generate optimized
    LangSAM
prompts. The optimized prompts must describe concrete visible objects,
    not
defects or abstract regions. Do not use words such as edge, artifact,
    shadow,
lighting, abnormal, inconsistent, region, or area.

Output strictly as JSON:
{
  "conclusion": "Good Enough" or "Needs Improvement",
  "evaluation_basis": "A concise explanation of mask completeness and
     errors.",
  "optimized_sam_prompts": []
}
"""
```