MADL: Towards Dependable Image Forgery Detection via Multi-Agent Forensic Reasoning

Dashan Qing, Xin Liao, *Senior Member, IEEE*, Chengyu Wang, and Jiaxin Chen

***Abstract*—The rapid advancement of generative models in creating realistic synthetic and manipulated images poses growing risks to visual-content authenticity. A practical forensic system must determine whether an image is authentic or forged, distinguish fully synthetic content from local manipulation, and locate tampered regions. Existing approaches typically rely on a single model for classification or localization, making it difficult to combine high-level semantic understanding with low-level manipulation traces when different evidence sources conflict. To address these challenges, we propose *MADL*, a multi-agent collaborative framework for image forgery detection and localization. MADL extends single-model discrimination into a hierarchical, evidence-driven collaborative adjudication process. It comprises three complementary agents: Agent A performs Qwen-based semantic verification and image-level three-class classification; Agent B extracts pixel-level manipulation evidence through dual-stream forensic analysis and candidate-mask ranking; and Agent C coordinates heterogeneous evidence through conflict-aware adjudication. Through structured evidence interaction and explicit agreement, suppression, override, and conflict-handling mechanisms, MADL combines semantic reasoning with manipulation-trace localization while keeping the decision process traceable. Experiments demonstrate competitive three-class classification and strong IoU-based tampered-region localization. Further ablations and decision-path audits validate the multi-agent collaboration and conflict-aware adjudication mechanisms.**

***Index Terms*—Image forgery detection, tampered-region localization, multi-agent reasoning, dependable forensics.**

I. Introduction

RECENT advances in generative artificial intelligence have substantially improved the realism and accessibility of synthetic and manipulated images. Modern generative models can synthesize visually convincing content, while text-guided editing systems can modify local objects, attributes, and semantic relations without introducing readily perceptible artifacts. For instance, X-Edit [1] enables instruction-driven local editing while preserving the visual coherence of unmodified regions. When such images are disseminated through social media or used for identity fraud, misinformation, evidence fabrication, and malicious content manipulation, they may mislead human observers and undermine trust in digital visual content [2]. Consequently, a practical image forensic system should answer three closely related questions: whether an image is authentic or forged, whether a forged image is fully synthetic or locally tampered, and where the manipulated regions are located.

Existing methods mainly follow three technical routes. Image-level detectors learn manipulation or generation traces from noise residuals and forensic representations, including rich-feature modeling [3], anomaly tracing [4], CNN-generated image detection [5], texture statistics [6], diffusion reconstruction error [7], universal visual features [8], and prompt-tuned semantic cues [9]. Robust forensic networks further suppress post-processing noise or separate it from manipulation traces [10], [11], while processing-chain analysis identifies coupled artifacts introduced by sequential global operations [12]. However, these image-level methods cannot simultaneously distinguish fully synthetic content from local tampering and identify the supporting region. Localization models estimate pixel-level masks from multi-view supervision [13], progressive correlation [14], hierarchical features [15], multi-clue analysis [16], diffusion priors [17], adaptive representations [18], progressive feedback [19], diffusion-based localization [20], or reconstruction cues [21]. Weak traces and post-processing may nevertheless produce incomplete or spurious responses. These limitations hinder unified three-class classification and localization.

Vision-language models introduce semantic reasoning into image forensics. LISA [22], FakeShield [23], TruthLens [24], VLForgery [25], SIDA [2], and AIGI-Holmes [26] support detection, localization, attribution, or explanation through multimodal representations. Yet semantic predictions and pixel-level traces may disagree: a plausible image-level judgment can lack spatial support, while a localization model may activate on irrelevant artifacts. A unified forensic system therefore requires an explicit mechanism for verifying and reconciling heterogeneous evidence rather than relying on a single prediction head or direct score fusion.

To this end, we propose *MADL*, a multi-agent collaborative framework for image forgery detection and localization. As illustrated in Fig. 1, Agent A performs Qwen-based semantic verification and three-class classification; Agent B extracts pixel-level evidence through dual-stream analysis, candidate-mask generation, and learned ranking; and Agent C conducts conflict-aware adjudication. Constrained evidence interfaces

This work is supported by the New Generation Artificial Intelligence–National Science and Technology Major Project (Grant No. 2025ZD0123601), the Science and Technology Innovation Program of Hunan Province (Grant No. 2025QK2008), the National Natural Science Foundation of China (Grant Nos. U22A2030 and 62402062), the National Key R&D Program of China (Grant No. 2024YFF0618800), Hunan Provincial Funds for Distinguished Young Scholars (Grant No. 2024JJ2025), the Natural Science Foundation of Hunan Province, China (Grant No. 2025JJ60415), and the Research Foundation of Education Bureau of Hunan Province, China (Grant No. 25B0221). (Corresponding author: Xin Liao.)

Dashan Qing, Xin Liao, and Chengyu Wang are with the College of Cyber Science and Technology, Hunan University, Changsha 410082, China (e-mail: qingds@hnu.edu.cn; xinliao@hnu.edu.cn; wangchengyu@hnu.edu.cn).

Jiaxin Chen is with the School of Computer Science and Technology, Changsha University of Science and Technology, Changsha 410076, China (e-mail: jxchen@csust.edu.cn).

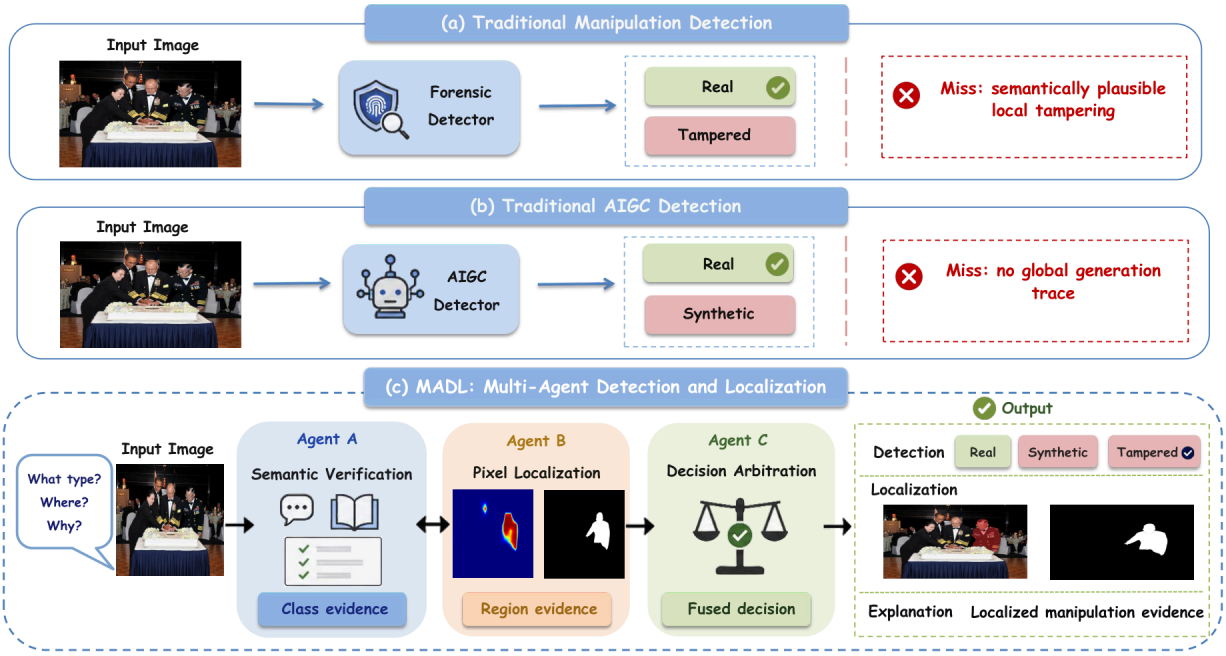


Fig. 1: Comparison between conventional image-forgery detection and MADL. MADL jointly performs three-class classification, tampered-region localization, and structured evidence tracing through three-agent collaboration.

and explicit agreement, suppression, override, and conflict states allow MADL to combine semantic and spatial evidence while retaining a traceable decision path.

Experiments show that MADL achieves competitive three-class classification and strong IoU-based tampered-region localization. Agent ablations identify the complementary roles of semantic verification, pixel-level evidence extraction, and conflict-aware adjudication, while decision-path audits quantify how conflicting evidence changes the final prediction.

Our main contributions are summarized as follows:

- We propose MADL, a multi-agent collaborative framework that reformulates image forgery analysis as hierarchical evidence adjudication. Unlike existing VLM-based methods that couple classification and localization within a single model or reasoning chain, MADL connects specialized agents through explicit heterogeneous-evidence interfaces and conflict-aware adjudication, while unifying the classification of authentic, fully synthetic, and locally tampered images with pixel-level localization.
- We develop three complementary agents connected by constrained heterogeneous-evidence interfaces. Agent A supplies semantic and classification evidence, Agent B extracts and selects pixel-level tampering evidence, and Agent C reconciles them through explicit agreement, suppression, override, and conflict states. This design preserves both the evidence source and the adjudication path of the final decision.
- Extensive experiments demonstrate competitive three-class classification and strong IoU-based localization. Agent ablations and decision-path audits further validate the complementary agent roles and conflict-aware adjudication mechanism.

II. Related Work

A. Image Forgery Detection and Localization

Image-level forgery detectors learn manipulation or generation artifacts from residual and semantic representations. Large-scale benchmarks such as GenImage [27] and DDL [28] support evaluation across diverse generation and manipulation settings. Learning Rich Features [3] and ManTra-Net [4] model local inconsistencies, whereas CNNSpot [5], LNP [29], GramNet [6], DIRE [7], UnivFD [8], and AntifakePrompt [9] target CNN- or diffusion-generated content and cross-generator generalization. Since compression, resizing, and sequential operations can obscure or couple manipulation artifacts, SNIS [10], PRest-Net [11], and feature-decoupled operator-chain identification [12] explicitly address post-processing robustness or processing-history evidence. Related research on perturbed-image representation [30] and screen-capture distortion modeling [31]–[33] further demonstrates the importance of retaining discriminative visual evidence under composite degradations. Although effective for authenticity analysis, operation discrimination, or robust representation, these methods do not establish spatial correspondence between a three-class prediction and its supporting region, which is necessary for distinguishing fully synthetic images from local tampering.

Localization methods estimate manipulated regions from boundary, noise, correlation, diffusion, or reconstruction cues. MVSS-Net [13], PSCC-Net [14], HiFi-Net [15], and TruFor [16] develop multi-scale or multi-clue localization, while DiffForensics [17], AdaIFL [18], ProFact [19], ForgDiffuser [20], Delocate [21], DCL-Net [34], and the dual-domain

multi-scale edge-guided network [35] introduce diffusion priors, adaptive representations, weak supervision, reconstruction, contrastive decoupling, or cross-domain edge guidance. Their emphasis remains low-level trace extraction; the consistency of incomplete or spurious masks with the image-level semantic judgment is rarely examined explicitly.

Multimodal methods introduce vision-language models (VLMs) into detection, localization, attribution, or explanation. LISA [22], FakeShield [23], TruthLens [24], VL-Forgery [25], SIDA [2], AIGI-Holmes [26], ForgeryGPT [36], ForgeryVCR [37], FOCA [38], and recent analyses of VLM-based forensics [39] extend forensic evidence from low-level artifacts to object relations and scene semantics. However, category prediction, localization, and semantic evidence are generally coordinated implicitly within one model or reasoning chain. MADL instead assigns semantic verification, image-level classification, and pixel-level analysis to specialized agents and explicitly reconciles their outputs for unified three-class classification and tampered-region localization.

B. Multi-Agent Collaboration for Image Forensics

Agent-based frameworks use role specialization and structured interaction to organize complex reasoning. PentestGPT [40] combines task decomposition with tools, while multi-agent debate [41] uses mutual critique. In image forensics, UniShield [42] routes inputs to domain-specific detectors and consolidates their outputs. These studies establish the value of agent organization and expert routing, but do not specify how semantic, classification, and pixel-level evidence should be reconciled when their predictions conflict.

MADL differs from routing-oriented systems in both objective and interaction design. Agent A supplies semantic and image-level classification evidence, Agent B extracts and selects pixel-level evidence, and Agent C adjudicates their outputs under explicit agreement or conflict states. This organization directly supports three-class classification and tampered-region localization while preserving the evidence source and decision path of the final result. Knowledge-distillation methods that transfer relative distributions or historical contrastive knowledge [43], [44] provide a complementary basis for reducing future large–small model collaboration costs, but do not themselves define forensic evidence interfaces or conflict-resolution states.

III. Methodology

A. Overview of MADL

The three-class classification of authentic, fully synthetic, and locally tampered images requires different but complementary forms of forensic evidence. Fully synthetic images are generally characterized by global generation patterns and semantic inconsistencies, whereas locally tampered images must be supported by spatially localized manipulation traces. A single model may produce an image-level prediction or a localization mask, but it may not consistently reconcile these heterogeneous cues when their predictions disagree. To address this issue, we propose MADL, which organizes semantic understanding, image-level classification, and pixel-level forensic analysis into a multi-agent collaborative process.

As illustrated in Fig. 2 and summarized in Table I, MADL consists of three agents with complementary responsibilities. Agent A performs Qwen-based semantic verification and produces image-level classification evidence. It analyzes the global image category and examines whether suspicious regions are semantically compatible with the observed image content. Agent B performs dual-stream forensic analysis to extract low-level manipulation traces, generate candidate masks, and select the candidate most strongly supported by the pixel-level evidence. Agent C receives the evidence produced by Agents A and B and performs conflict-aware adjudication to determine the final image category and localization result.

The collaborative process begins with an initial analysis by Agent A, followed by pixel-level verification by Agent B and a candidate-anchored review by Agent A. Agent C then coordinates the resulting evidence:

$$\begin{aligned}(E_A, E_B) &= \mathcal{F}_{AB}(x),\\ (\hat{y}, \hat{M}, \mathcal{D}) &= C(E_A, E_B),\end{aligned} \tag{1}$$

where $\mathcal{F}_{AB}$ denotes the ordered evidence-construction process between Agents A and B: Agent A first produces initial semantic and spatial-prior evidence, Agent B converts the priors and its bottom-up forensic responses into pixel-supported candidates, and Agent A subsequently reviews those candidates. The resulting packets E_A and E_B contain semantic/classification evidence and pixel-level localization evidence, respectively. The outputs $\hat{y}$, $\hat{M}$, and $\mathcal{D}$ denote the final three-class label, the localization mask retained for a locally tampered image, and the collaborative decision path.

The Agent identifiers indicate functional responsibilities rather than a single-pass execution order. During inference, Agent A first produces a global three-class hypothesis and weak spatial priors. Agent B converts these priors, together with its bottom-up anomaly responses, into pixel-supported mask candidates. Agent A then reviews the resulting candidates before Agent C examines whether the semantic, classification, and pixel-level evidence are mutually consistent. When the evidence sources agree, their predictions are retained. When they disagree, Agent C applies the agreement, suppression, override, or conflict-handling state according to the available evidence. This collaboration also distinguishes local manipulation from full-image synthesis. A locally tampered prediction requires a valid mask and sufficient pixel-level support. If reliable local manipulation evidence is unavailable, Agent C uses the image-level evidence from Agent A to distinguish an authentic image from a fully synthetic image. MADL therefore integrates three-class image classification and tampered-region localization without treating a fully synthetic image as a whole-image tampering mask.

B. Agent A: Semantic Verification and Image-Level Classification

Pixel-level forensic models can capture local noise patterns, boundary variations, and imaging inconsistencies, but low-level traces alone may be insufficient to determine whether a

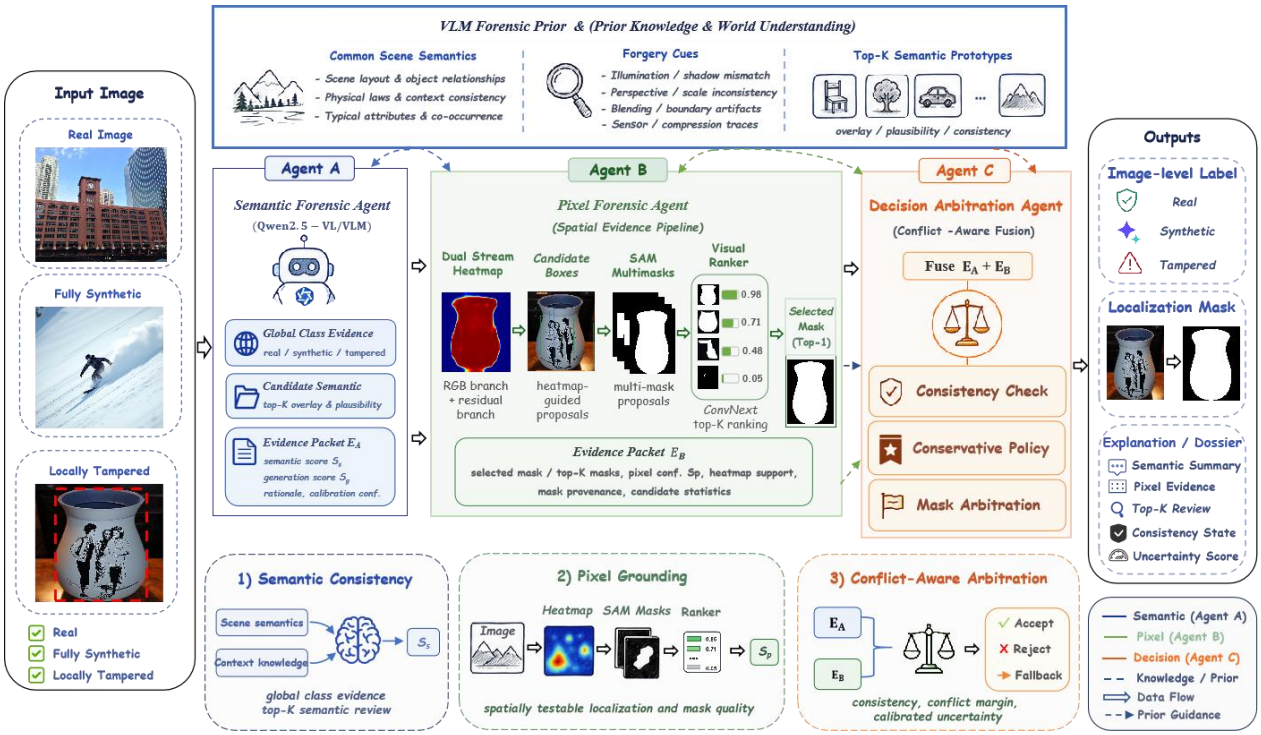


Fig. 2: Overall architecture of MADL. Agent A performs Qwen-based semantic verification and produces image-level classification evidence. Agent B extracts pixel-level manipulation evidence and generates the corresponding candidate masks. Agent C coordinates the heterogeneous evidence and produces the final three-class label and, when local tampering is detected, the localization mask.

TABLE I: Functional Roles and Evidence Exchange of the Three Agents in MADL.

Agent	Responsibility	Input	Output
Agent A	Qwen-based semantic verification and image-level classification	Image x and pixel-supported candidates from Agent B	Initial class and spatial-prior evidence $E_A^{(0)}$, followed by candidate-anchored semantic evidence E_A
Agent B	Pixel-level forensic analysis, candidate-mask generation, and learned mask ranking	Image x and weak spatial priors from Agent A	Anomaly response, candidate masks, selected mask, pixel confidence, and evidence packet E_B
Agent C	Coordination of heterogeneous evidence and conflict-aware adjudication	Evidence packets E_A and E_B	Final three-class label, conditional localization mask, and collaborative decision path $\mathcal{D}$

suspicious region is anomalous in terms of object relations, scene structure, or generation semantics. This limitation is particularly relevant to fully synthetic images, whose forensic evidence is generally distributed across the entire image rather than associated with a specific local region. MADL therefore introduces Agent A, which employs a Qwen vision-language model adapted to the forensic task and combines its semantic analysis with an image-level classification head to distinguish authentic, fully synthetic, and locally tampered images.

Agent A operates in two stages: initial forensic analysis and candidate-anchored review. During the initial stage, Agent A receives the input image x and uses Qwen to jointly analyze the image category, evidence scope, and suspicious regions. Instead of requesting an unconstrained natural-language explanation, forensic supervised fine-tuning (SFT) constrains Qwen to produce a structured result containing a three-class hypothesis, an evidence scope, and candidate regions. For a locally tampered image, the evidence scope is local and the model provides the corresponding candidate boxes. For a fully synthetic image, the evidence scope is global, without treating the entire image as a local manipulation region.

The initial evidence produced by Agent A is represented as

$$E_A^{(0)} = \mathcal{A}_{\text{init}}(x) = \{\mathbf{p}_A, r_A, \mathcal{R}_A\}, \tag{2}$$

where $\mathbf{p}_A = [p_A^r, p_A^s, p_A^t] = \text{softmax}(\mathbf{g}_A(x))$ denotes the normalized three-class evidence produced from the logits $\mathbf{g}_A(x)$ of the image-level classification head adapted through SFT. No post-hoc probability-calibration method, such as temperature scaling, is applied to these scores. The term r_A denotes Qwen's structured semantic assessment, and $\mathcal{R}_A = \{R_i\}_{i=1}^{N_A}$ denotes the candidate-region set generated by Qwen. Each candidate region contains normalized coordinates, evidence scope, and the corresponding confidence information. These regions serve only as spatial priors for the pixel-level analysis performed

by Agent B and are not directly treated as final manipulation masks.

After Agent B performs pixel-level forensic analysis, candidate-mask generation, and mask ranking, Agent A further reviews the candidates supported by the pixel-level evidence. Specifically, Agent A jointly examines the input image and the visualized candidate regions to assess boundary transitions, illumination consistency, geometric plausibility, object–scene compatibility, and global generation characteristics. This process does not regenerate the manipulation mask. Instead, it evaluates whether the spatial evidence produced by Agent B is semantically compatible with the observable image content:

$$E_A = \mathcal{A}_{\text{review}}\left(x, E_B; E_A^{(0)}\right). \tag{3}$$

The resulting evidence packet E_A retains the initial three-class evidence and further includes the candidate-level semantic-consistency result, local semantic anomaly score, and global synthesis evidence. Agent C subsequently coordinates E_A with the pixel-level evidence packet E_B to obtain the final forensic decision.

Agent A implements a two-stage reasoning process that connects global semantic analysis with candidate-anchored verification. Qwen analyzes image content, object relations, and generation semantics and provides suspicious-region priors for local tampering analysis, while the classification head supplies the normalized three-class evidence in $\mathbf{p}_A$. After Agent B extracts pixel-level forensic evidence, Agent A further evaluates the plausibility of the candidate regions in terms of boundary transitions, illumination consistency, geometric relationships, and scene context. Through this process, Agent A provides high-level semantic evidence for three-class classification and semantically verifies candidates supported by pixel-level evidence. Agent A neither generates the final localization mask nor independently determines the final class; Agent C adjudicates its evidence together with the pixel-level evidence provided by Agent B.

C. Agent B: Pixel-Level Evidence Extraction and Tampering Localization

Agent A can identify semantically suspicious regions, but its candidate boxes provide only coarse spatial priors and cannot directly represent precise manipulation boundaries. Moreover, the masks produced by a general-purpose segmentation model such as the Segment Anything Model (SAM) [45] primarily reflect object integrity, and their confidence scores do not necessarily indicate whether a mask is supported by forensic traces. MADL therefore introduces Agent B, which integrates dual-stream pixel forensics, candidate-mask generation, and learned mask ranking into a unified pixel-level forensic process. Agent B receives the input image and the weak spatial priors supplied by Agent A, and outputs the pixel-level anomaly response, candidate masks, selected localization mask, and corresponding pixel confidence.

1) Dual-Stream Pixel Evidence Extraction: Agent B first extracts complementary forensic features from the RGB image and its high-frequency residual representation. The RGB stream preserves image content, texture, and boundary structures, whereas the residual stream employs spatial rich model (SRM)-style high-pass filters [46] to suppress dominant semantic content and emphasize local noise distributions, resampling traces, and imaging inconsistencies. The two inputs are encoded by parallel ConvNeXt-Tiny [47] backbones, followed by shallow cross-attention for feature interaction. The fused representation is then processed by a classification head and a heatmap head to produce pixel-side three-class evidence and a pixel-level anomaly map, respectively.

Let x_r denote the RGB input and x_n denote the corresponding high-frequency residual. The dual-stream network produces

$$(\mathbf{o}_B, H_B) = \mathcal{P}_\theta(x_r, x_n), \tag{4}$$

where $\mathbf{o}_B \in \mathbb{R}^3$ denotes the pixel-side logits for the authentic, fully synthetic, and locally tampered categories, and $H_B \in [0,1]^{H\times W}$ denotes the pixel-level anomaly heatmap. The local-tampering evidence and the pixel-side synthesis evidence are obtained from the posterior probabilities:

$$\begin{aligned} S_B^t &= \text{softmax}(\mathbf{o}_B)_{\text{tampered}}, \\ S_B^s &= \text{softmax}(\mathbf{o}_B)_{\text{synthetic}}. \end{aligned} \tag{5}$$

Here, S_B^t measures the support for local manipulation, while S_B^s represents the global synthesis evidence captured by the pixel stream. The anomaly heatmap H_B provides the spatial distribution of the detected manipulation traces.

2) Candidate-Mask Generation: Agent B constructs initial candidate regions from two complementary sources: connected anomaly responses in H_B and the weak spatial priors supplied by Agent A. For each candidate box b_j, SAM generates multiple mask hypotheses through box scaling, multi-mask inference, and optional center-point prompting for small regions:

$$C_j = \left\{ (M_{j,k}, s_{j,k}^{\text{sam}}, g_{j,k}) \right\}_{k=1}^{K_j}, \tag{6}$$

where $M_{j,k}$ is the k-th candidate mask, $s_{j,k}^{\text{sam}}$ is the SAM confidence, and $g_{j,k}$ records auxiliary information such as proposal scale, mask area, boundary contact, and heatmap support.

Agent B first removes low-confidence or geometrically implausible candidates and retains a limited number of valid masks for each proposal. The remaining SAM masks are pooled with heatmap-direct candidates and box-fallback hypotheses to form the final candidate set. This stage improves candidate coverage but does not directly determine which mask best represents the manipulated region.

3) Learned Mask Ranking: The confidence produced by SAM mainly measures segmentation quality rather than forensic relevance. Agent B therefore employs a learned visual ranker to identify the candidate most strongly supported by the pixel-level forensic evidence.

For a candidate mask M_k, let B_k denote its tight bounding box with additional boundary padding. The corresponding RGB crop, candidate mask, and anomaly heatmap are concatenated into a five-channel input:

$$T_k = \text{Resize}_{160\times 160}\left([x_{B_k}, M_{k,B_k}, H_{B,B_k}]\right). \tag{7}$$

A feature vector g_k further describes the proposal geometry, mask occupancy, boundary contact, SAM and candidate confidence, and anomaly-heatmap support. The visual ranker

combines the local visual representation and these statistical features to predict the forensic-support score:

$$q_k = \text{sigmoid}\left(f_\theta(T_k, g_k)\right), \quad k^\star = \arg\max_k q_k, \qquad \hat{M}_B = M_{k^\star}. \tag{8}$$

During training, the ranker is supervised using the IoU between each candidate mask and the ground-truth manipulation mask. It therefore learns which candidate is most consistent with the actual manipulated region rather than merely learning general object-segmentation quality. During inference, ground-truth masks, candidate IoU values, and oracle information are unavailable; the final mask is selected entirely from the input image, anomaly heatmap, and candidate statistics.

4) Training Objective and Evidence Output: The dual-stream component of Agent B is jointly optimized using image-level classification and pixel-level localization losses:

$$\mathcal{L}_B = \alpha\mathcal{L}_{\text{cls}} + \beta\mathcal{L}_{\text{bce}} + \gamma\mathcal{L}_{\text{dice}}, \tag{9}$$

where $\mathcal{L}_{\text{cls}}$ provides three-class classification supervision, while $\mathcal{L}_{\text{bce}}$ and $\mathcal{L}_{\text{dice}}$ constrain the predicted anomaly heatmap using the ground-truth manipulation mask. The frozen Agent B model uses $(\alpha, \beta, \gamma) = (0.2, 2.0, 4.0)$, as recorded in its training configuration. The classification term is down-weighted because the principal function of this branch is to supply spatial evidence rather than replace Agent A's image-level decision. BCE retains dense pixelwise supervision, while the larger Dice weight counteracts foreground–background imbalance and emphasizes overlap for the often sparse manipulated region. These weights were selected before the formal evaluation, kept unchanged for all reported experiments, and were not tuned on the evaluation partition. The mask losses are applied only to locally tampered images with pixel-level annotations. Authentic and fully synthetic images contribute to classification supervision without treating the entire image as a manipulated region. This strategy encourages Agent B to focus on local manipulation traces and prevents fully synthetic images from being modeled as whole-image tampering masks.

The resulting pixel-level evidence packet is represented as

$$E_B = \left\{\mathbf{o}_B, H_B, C, \hat{M}_B, S_B^t, S_B^s\right\}. \tag{10}$$

Agent B does not independently determine the final image category. Its evidence is returned to Agent A for candidate-anchored semantic review and subsequently submitted to Agent C for collaborative adjudication.

D. Agent C: Conflict-Aware Collaborative Adjudication

Agent C converts the heterogeneous evidence produced by Agents A and B into the final forensic output. It neither extracts new visual features nor generates candidate masks. Instead, it determines whether the semantic and pixel-level evidence should be accepted, suppressed, or overridden under explicit decision conditions. Its inputs include Agent A's normalized class evidence $\mathbf{p}_A$, Qwen semantic evidence, and candidate-review result, together with Agent B's local-tampering score S_B^t, selected mask M_B, and proposal provenance ρ_B. The magnitude of cross-agent disagreement is recorded as

$$\Delta = \left|S_B^t - S_A^t\right|, \tag{11}$$

which serves as a diagnostic quantity rather than an additional prediction score.

To make these states deterministic, let $v_B = \mathbb{I}[M_B \neq \varnothing]$ denote mask validity, where a mask is valid only if the selected proposal contains a nonempty stored mask. The initial local-tampering indicator is

$$z_B = \mathbb{I}\left[v_B = 1 \ \wedge \ S_B^t \geq \tau_B\right]. \tag{12}$$

We further define strong authentic evidence from Agent A, a Qwen-supported local-tampering predicate, and a weak pixel proposal as

$$\begin{aligned} a_r &= \mathbb{I}[\hat{c}_A = r \wedge p_A^r \geq \tau_A^r \wedge p_A^t \leq \bar{\tau}_A^t], \\ q_t &= \mathbb{I}[\hat{c}_Q = t \wedge c_Q \geq \tau_Q \wedge \max(S_Q^g, S_Q^m) \geq \tau_Q^t], \\ w_B &= \mathbb{I}[S_B^t \leq \tau_B^w \wedge \rho_B = \text{layout}], \end{aligned} \tag{13}$$

where $\hat{c}_A = \arg\max_c p_A^c$, $\hat{c}_Q$ and c_Q are Qwen's structured class hypothesis and semantic confidence, and S_Q^g and S_Q^m are its global and candidate-anchored tampering scores. Suppression is activated iff $\sigma = z_B a_r w_B (1 - q_t) = 1$. A supported override is activated iff

$$\omega = \mathbb{I}[z_B = 0 \wedge v_B = 1 \wedge p_A^t \geq \tau_A^t \wedge p_A^r \leq \bar{\tau}_A^r] = 1. \tag{14}$$

The adjudicated local state is therefore

$$z_C = \begin{cases} 0, & z_B = 1 \wedge \sigma = 1, \\ 1, & z_B = 1 \wedge \sigma = 0, \\ 1, & z_B = 0 \wedge \omega = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{15}$$

Thus, agreement retains a supported pixel decision; suppression rejects only a weak layout-derived proposal under strong authentic evidence without Qwen support; override requires both Agent A tampering evidence and a valid pixel mask; and the remaining disagreement is recorded as conflict. Table III lists every threshold used by these predicates. The diagnostic disagreement Δ is retained in the trace but is not thresholded and does not alter the decision. The final three-class label and localization mask are then obtained by

$$\begin{aligned} c_A^\star &= \arg\max_{c \in \{\text{real}, \text{synthetic}\}} p_A^c, \\ (\hat{y}, \hat{M}) &= \begin{cases} (\text{tampered}, M_B), & z_C = 1, \\ (c_A^\star, \varnothing), & z_C = 0. \end{cases} \end{aligned} \tag{16}$$

For each input, Agent C stores a structured decision trace

$$\mathcal{D} = \{s_C, \mathbf{p}_A, S_A^t, S_B^t, \Delta, M_B, d_C\}, \tag{17}$$

where s_C denotes the agreement, suppression, override, or conflict state, and d_C records the evidence source responsible for the final decision. Therefore, the role of Agent C is not simple score fusion, but explicit adjudication over heterogeneous evidence with a traceable decision state.

IV. Experiments

We evaluate MADL through three-class image detection, tampered-region localization, robustness under post-processing perturbations, agent-level ablation, localization-component ablation, decision-path analysis, and qualitative visualization. These experiments assess both the overall forensic performance of MADL and the roles of semantic analysis, pixel-level evidence extraction, and conflict-aware collaborative adjudication.

A. Experimental Setup

Datasets. SID-Set [2] contains 300,000 images evenly divided among authentic, fully synthetic, and locally tampered classes, comprising 210,000 training images (70,000 per class), 30,000 validation images (10,000 per class), and 60,000 test images (20,000 per class). Authentic images originate from OpenImages V7; fully synthetic images are generated with FLUX using Flickr30k and COCO content; and locally tampered images are produced by object replacement and partial attribute editing with object masks and latent-diffusion editing. Pixel-level ground-truth masks are provided for locally tampered images, and localization is evaluated on this subset.

Implementation Details. MADL is implemented in Python 3.10 and PyTorch and evaluated on a single NVIDIA RTX 5090 GPU. Agent A adapts Qwen2.5-VL-7B-Instruct [48] through SFT with low-rank adaptation (LoRA) [49]. Agent B employs a dual-stream ConvNeXt-Tiny [47] with SRM residual features [46], while SAM ViT-H generates candidate masks. OpenCV and Pillow are used for image preprocessing. The Agent C thresholds are selected by grid search on a fixed 20% calibration subset that is disjoint from the held-out evaluation subset. Candidate settings are first required to satisfy Real FPR $\leq$ 20%, Tampered Recall $\geq$ 85%, and IoU $\geq$ 65%; feasible settings are then ranked by Accuracy, IoU, Tampered Recall, and Real FPR. The selected values in Table III are frozen for every reported test and perturbation condition.

Evaluation Metrics. Three-class detection is evaluated using class-wise Accuracy and F1, together with overall Accuracy and Macro-F1. Tampered-region localization is evaluated using the pixel-level area under the receiver operating characteristic curve (AUC), F1, and IoU. Hereafter, AUC and F1 in localization tables refer to pixel-level metrics. The real-image false-positive rate (Real FPR) and tampered-image recall (Tampered Recall) are additionally reported in the agent ablation and decision-path audit.

Comparison Methods. For image-level detection, MADL is compared with AntifakePrompt [9], CNNSpot [5], Gram-Net [6], UnivFD [8], SIDA-7B [2], and SIDA-13B [2]. Following the SID-Set benchmark protocol, AntifakePrompt, CNNSpot, Gram-Net, and UnivFD are evaluated directly with the authors' pretrained models, without SID-Set retraining, an additional classification head, or binary-to-three-class score mapping. Their rows report native real-versus-forged decisions conditioned on the authentic, fully synthetic, and locally tampered subsets; they therefore serve as category-wise transfer references rather than native three-class predictors. SIDA-7B, SIDA-13B, and MADL directly produce the three labels. For tampered-region localization, the comparison includes MVSS-Net [13], HiFi-Net [15], PSCC-Net [14], LISA-7B [22], and SIDA-7B [2].

B. Three-Class Detection Performance

We first evaluate whether MADL can jointly distinguish authentic, fully synthetic, and locally tampered images on SID-Set. This setting differs from conventional binary detection because the model must not only identify manipulated content, but also determine whether the forensic evidence is distributed globally across a fully synthetic image or concentrated within a locally tampered region.

As shown in Table II, MADL achieves an overall Accuracy of 91.83% and a Macro-F1 of 91.77%. Its class-wise Accuracy is 80.71% for authentic images, 99.02% for fully synthetic images, and 95.66% for locally tampered images. The conventional binary detectors provide category-conditioned transfer references, whereas MADL and SIDA directly distinguish the three forensic categories. The localization evaluation and agent ablation below further examine the spatial evidence and the respective contribution of each agent.

C. Localization Results

Image-level detection is insufficient for forensic use because practical analysis also requires identifying the manipulated regions. We therefore compare MADL with representative localization baselines, including MVSS-Net, HiFi-Net, PSCC-Net, LISA-7B, and SIDA-7B.

As shown in Table IV, MADL achieves 95.74% AUC, 78.32% F1, and 70.33% IoU, with a 95% confidence interval of 69.75–70.94 for IoU. Compared with SIDA-7B, MADL improves AUC, F1, and IoU by 8.44, 4.42, and 26.53 percentage points, respectively. Agent B generates the final masks

TABLE II: Three-Class Image Detection Performance on SID-Set.

Method	Authentic		Fully Synthetic		Locally Tampered		Overall	
	Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	Macro-F1
AntifakePrompt [9]	64.80	78.60	93.80	96.80	30.80	47.20	63.10	69.30
CNNSpot [5]	79.80	88.70	39.50	56.60	6.90	12.90	42.10	69.60
Gram-Net [6]	70.10	82.40	93.50	96.60	0.80	1.60	55.00	58.00
UnivFD [8]	68.00	67.40	62.10	87.50	64.00	85.30	64.00	85.30
SIDA-7B [2]	89.10	91.00	98.70	98.60	91.20	91.00	93.50	**93.50**
SIDA-13B [2]	**89.60**	**91.10**	98.50	98.70	92.90	**91.20**	**93.60**	**93.50**
MADL (Ours)	80.71	86.84	**99.02**	**99.44**	**95.66**	89.02	91.83	91.77

TABLE III: Frozen Thresholds and Deterministic Conditions Used by Agent C.

Role	Symbol/condition	Value
Local decision	τ_B	0.50
Weak pixel evidence	τ_B^w	0.80
Strong authentic evidence	$\tau_A^r, \bar{\tau}_A^t$	0.93, 0.04
Qwen local support	τ_Q, τ_Q^t	0.45, 0.05
Tampered evidence	$\tau_A^t, \bar{\tau}_A^r$	0.05, 0.95
Suppressible proposal	ρ_B	layout
Valid mask	$M_B \neq \varnothing$	deterministic

TABLE IV: Pixel-Level Tampering Localization Performance on SID-Set.

Method	AUC	F1	IoU
MVSS-Net [13]	48.90	31.60	23.70
HiFi-Net [15]	64.00	45.90	21.10
PSCC-Net [14]	82.10	71.30	35.70
LISA-7B [22]	78.40	69.10	32.50
SIDA-7B [2]	87.30	73.90	43.80
MADL (Ours)	**95.74**	**78.32**	**70.33**

TABLE V: Performance Across Different Tampered-Region Scales.

Tampered area ratio	Accuracy	F1	IoU	Success@0.5
Small (< 5%)	94.03	66.21	57.53	66.55
Medium (5–20%)	96.43	83.32	74.75	88.21
Large (> 20%)	**96.76**	**88.12**	**81.37**	**92.49**

through heatmap-guided proposal generation, SAM candidate construction, candidate pruning, and learned mask ranking, while Agent C only determines whether the selected mask is retained in the final output.

1) Effect of Tampered-Region Scale: To examine whether localization performance depends on the spatial extent of manipulation, locally tampered images are divided according to the ratio between the ground-truth tampered area and the image area. Small, medium, and large manipulations correspond to area ratios below 5%, between 5% and 20%, and above 20%, respectively.

Table V shows that localization improves consistently as the manipulated region becomes larger. IoU increases from 57.53% for small regions to 81.37% for large regions. The larger gap in localization metrics than in classification Accuracy indicates that small manipulations primarily challenge spatial recovery rather than image-level recognition.

2) Effect of Boundary Complexity: Beyond area, irregular boundaries increase the difficulty of recovering precise mask geometry. We quantify ground-truth boundary complexity using the squared external-contour perimeter normalized by foreground area, and divide the samples into three equal-frequency groups. As shown in Table VI, IoU decreases from 78.92% to 61.65% as boundary complexity increases, while classification Accuracy remains comparatively stable. This result indicates that complex contours mainly affect boundary alignment.

3) Effect of Region Multiplicity: Local manipulations may form either a single connected region or multiple spatially separated regions. Significant connected components are identified with 8-connectivity after excluding components smaller than 0.1% of the foreground area. Multiple-region samples obtain 59.38% IoU, compared with 73.07% for single-region samples. The decrease suggests that proposal coverage and mask selection become more difficult when the forensic evidence is spatially fragmented.

TABLE VI: Localization Performance Under Different Mask Structures.

Factor	Group	F1	IoU	Success@0.5
Boundary complexity	Low	**84.83**	**78.92**	**88.42**
	Medium	78.62	70.42	83.23
	High	71.97	61.65	73.20
Region multiplicity	Single	**80.68**	**73.07**	**84.63**
	Multiple	69.68	59.38	69.59

Fig. 3 complements the aggregate metrics by comparing spatial predictions from MVSS-Net, HiFi-Net, SIDA-7B, and MADL on identical inputs. The cases cover diverse manipulation scales, structures, and scene contents, allowing localization extent, boundary integrity, and missed responses to be inspected directly. Because the compared methods are learned under different training settings, this figure characterizes their qualitative localization behavior and does not replace the quantitative evaluation in Table IV.

D. Localization-Branch Robustness

Images distributed through online platforms often undergo recompression, resizing, noise contamination, format conversion, or blur. Table VII evaluates the frozen Agent B localization branch under JPEG compression, image resizing, Gaussian noise, WebP conversion, and Gaussian blur. All perturbation conditions use the same fixed sample manifest, model parameters, mask-score threshold, and candidate-selection procedure.

Across JPEG quality factors from 50 to 80, localization F1 remains between 76.71% and 78.39%, while IoU remains between 68.57% and 70.33%. Downscaling to 0.5 or 0.75 and upscaling to 1.5 likewise preserve IoU between 69.45% and 71.22%; the 1.5x condition attains 78.09% localization F1 and 69.89% IoU. Gaussian noise has a more pronounced effect as its intensity increases: localization F1 decreases from 73.42% to 64.45%, while IoU decreases from 65.47% to 56.29%. WebP conversion and Gaussian blur produce intermediate localization results. The variation in image-level Accuracy, Real FPR, and Tampered Recall further indicates that post-processing affects the mask-score decision boundary differently across authentic and tampered images. These results show tolerance to common recompression and scale changes, while also identifying noise-induced trace attenuation as a clear limitation.

E. Ablation Study

Table VIII progressively introduces the three agents on the same real-versus-tampered evidence-adjudication subset. All configurations reuse identical pixel predictions and sample order. Agent B provides the pixel-level forensic baseline; Agents

TABLE VII: Localization-Branch Robustness Under Common Image Perturbations.

Perturbation	Accuracy	Cls. F1	Real FPR	Tampered Recall	Loc. F1	IoU
JPEG 50	69.40	74.20	49.20	88.00	77.97	70.02
JPEG 60	70.60	74.79	46.00	87.20	77.71	69.61
JPEG 70	71.40	75.72	46.40	89.20	78.39	70.33
JPEG 80	71.80	76.06	46.00	**89.60**	76.71	68.57
Resize 0.5	**78.20**	80.15	31.60	88.00	**79.29**	**71.22**
Resize 0.75	74.60	77.68	39.20	88.40	77.63	69.45
Resize 1.5	74.60	77.91	40.40	**89.60**	78.09	69.89
Gaussian 5	74.80	76.67	33.20	82.80	73.42	65.47
Gaussian 10	72.40	72.62	**28.40**	73.20	64.45	56.29
WebP 80	68.20	72.44	47.20	83.60	75.28	67.27
Blur 3×3	**78.20**	**80.36**	32.80	89.20	76.01	67.86

TABLE VIII: Progressive Agent Ablation on the SID-Set Real-Versus-Tampered Evidence Subset.

Configuration	Acc.↑	Macro-F1↑	Real FPR↓	Tamp. Rec.↑
Agent B	74.57	74.04	39.69	**88.80**
Agents A+B	77.60	77.44	30.70	85.88
MADL (Agents A+B+C)	**83.75**	**83.90**	**17.62**	85.61

A+B additionally introduce SFT-adapted Qwen semantic verification with simple evidence fusion; and the complete MADL configuration further activates the conflict-aware adjudication of Agent C.

Compared with Agent B, introducing Agent A increases Accuracy from 74.57% to 77.60% and Macro-F1 from 74.04% to 77.44%, while reducing Real FPR from 39.69% to 30.70%. This result shows that the VLM contributes high-level semantic verification that complements, rather than replaces, pixel-level manipulation analysis. Agent C further increases Accuracy and Macro-F1 to 83.75% and 83.90%, respectively, and reduces Real FPR to 17.62%. The progressive improvements demonstrate that semantic verification, pixel-level evidence, and conflict-aware adjudication provide complementary contributions within MADL.

Agent A: VLM Backbone Replacement. To examine whether the semantic-verification capability of Agent A depends on a specific VLM, we replace its backbone while retaining the remaining MADL components. As shown in Table IX, the fine-tuned Qwen2.5-VL backbone achieves the strongest overall Accuracy, Tampered Recall, and IoU among the evaluated variants. Compared with the base Qwen2.5-VL model, task-specific fine-tuning substantially improves the forensic capability of Agent A, supporting the use of the adapted Qwen2.5-VL as the default VLM backbone in MADL.

Agent A: Semantic-Verification Analysis. To examine the semantic path independently of the simple fusion used in Ta-

Fig. 3: Qualitative comparison of same-sample tampered-region localization on SID-Set.

TABLE IX: VLM Backbone Replacement for Agent A.

Agent A Backbone	Fine-tuned	Overall Accuracy↑	Real FPR↓	Tampered Recall↑	IoU↑
Qwen2.5-VL-7B-Instruct	No	58.55	53.27	70.36	51.25
LLaVA-OneVision-7B	No	69.31	**4.53**	43.14	30.90
InternVL3-8B	No	48.79	50.97	48.55	34.99
Qwen2.5-VL-FT	**Yes**	**85.21**	17.46	**87.87**	**65.40**

TABLE X: Five-Fold Linear-Probe Evaluation of Image-Token Representations.

Model	Authentic Recall	Synthetic Recall	Tampered Recall	Balanced Acc.	Macro-F1
Base Qwen	76.67	83.33	63.33	74.44	73.98
SIDA-7B	70.00	86.67	50.00	68.89	67.97
SFT-adapted Qwen	**83.33**	**100.00**	**76.67**	**86.67**	**86.24**

TABLE XI: Localization-Component Ablation for Agent B.

Configuration	Candidate prior	Mask selection	AUC	F1	IoU	Proposal Recall@0.5
Heatmap direct	DualStream heatmap	Fixed threshold	**96.14**	73.12	63.30	–
SAM + learned ranking	DualStream heatmap	Learned ranking	**96.14**	69.58	60.98	**88.00**
Full Agent B	Heatmap-guided candidates	Candidate refinement	**96.14**	**77.87**	**70.15**	**88.00**

TABLE XII: Quantitative Audit of Conflict-Aware Decision States.

Decision state	Accuracy before C	Accuracy after C	Corrected (%)	Worsened (%)	ΔReal FPR	ΔTampered Recall
Agreement	84.87	84.87	0.00	0.00	0.00	0.00
Suppression	13.36	86.64	86.64	13.36	-100.00	-100.00
Override	10.53	89.47	89.47	10.53	100.00	100.00
Conflict	82.68	82.19	0.00	0.49	2.76	0.00
Overall	74.57	83.75	11.14	1.96	-21.58	-3.18

ble VIII, we replay Agent A on the same frozen pixel evidence. Agent A changes 957 initial pixel-supported decisions, correcting 715 and worsening 242; the paired change is significant under an exact McNemar test ($p < 0.001$). These results indicate that Agent A mainly suppresses visually plausible false alarms by checking whether the localized evidence is semantically consistent with the image content.

Table X further evaluates the image-token representations of Base Qwen, SFT-adapted Qwen, and SIDA-7B on the same fixed class-balanced sample of 90 inputs (30 per class). For each model, final-layer image tokens are mean-pooled and evaluated using an identical five-fold stratified linear-probe protocol; feature standardization and multinomial logistic-regression fitting are restricted to the training portion of each fold. SFT-adapted Qwen achieves the highest recall in all three forensic categories and improves Balanced Accuracy from 74.44% to 86.67% and Macro-F1 from 73.98% to 86.24% over Base Qwen. Its paired out-of-fold predictions correct 13 Base-Qwen errors while introducing two new errors (exact McNemar test, $p = 0.0074$). This controlled diagnostic indicates that task adaptation makes Agent A's semantic representation more linearly accessible; it complements, rather than replaces, the end-to-end VLM comparison in Table IX.

Agent B: Localization-Component Analysis. Table XI isolates the stages that directly affect mask geometry on the fixed component-ablation manifest. The AUC column evaluates the shared continuous dual-stream heatmap and is consequently identical across the three configurations; F1 and IoU evaluate the final binary mask produced by each configuration. Heatmap-direct thresholding provides a strong initial mask, while the complete Agent B improves F1 and IoU through candidate refinement and learned selection.

F. Decision-Path Audit

Agent C: Conflict-Aware Adjudication Analysis. The decision-path experiment compares the prediction before and after Agent C on identical frozen-evidence records. The recorded decision source is grouped into agreement, suppression, override, and residual conflict states. Table XII reports the accuracy before and after adjudication, together with the proportions of corrected and worsened decisions. The state-wise values are conditional rates within each decision state and do not represent the prevalence of that state. Agent C raises overall Accuracy from 74.57% to 83.75%; 11.14% of all initial decisions are corrected and 1.96% are worsened. The paired change is significant under an exact McNemar test ($p < 0.001$).

G. Qualitative Results

Fig. 4 presents complementary qualitative views of MADL. Fig. 4(a) contains representative authentic, fully synthetic, and locally tampered images from diverse scenes. These examples illustrate the unified output space of the framework: authentic and fully synthetic inputs receive image-level labels, whereas locally tampered inputs additionally invoke the spatial evidence path. The visual diversity across objects, people, outdoor scenes, and generated content also shows that the three-class prediction is not tied to a single semantic category.

The three rows in Fig. 4(a) also expose the distinction that motivates the task formulation. The authentic row contains

(a) Three-class predictions. (b) Localization and explanations.

Fig. 4: Qualitative results of MADL: (a) representative predictions for authentic, fully synthetic, and locally tampered images; (b) evidence traces for locally tampered images, including the input, heatmap, predicted mask, overlay, and structured explanation.

natural scenes, animals, and people without an associated local mask. The fully synthetic row includes both human-centered content and outdoor objects; although individual objects may appear plausible, the decision concerns the image as a globally generated sample rather than a locally edited region. By contrast, the locally tampered row contains otherwise coherent scenes in which only a bounded object or object part is manipulated. Treating these cases as a single binary fake class would hide the difference between global generation evidence and spatially bounded tampering evidence. The displayed output therefore reflects both the category of forgery and whether a localization result is forensically meaningful.

Fig. 4(b) further traces four locally tampered cases through the input, anomaly heatmap, predicted mask, mask overlay, and structured explanation. The heatmap first identifies a suspicious response around the manipulated object; the selected mask then converts this response into an explicit spatial hypothesis, and the overlay makes its correspondence with the image directly inspectable. The final text summarizes the localized inconsistency used in the decision. These examples are intended as evidence traces rather than a standalone evaluation of explanation quality: they show how a prediction can be followed from low-level response to region selection and semantic description, while the quantitative localization results are reported separately in Tables IV–VI.

The four cases cover different region sizes and geometries. The mouse and keyboard examples contain compact object-level changes with comparatively regular boundaries, whereas the train example occupies a larger, irregular region and the bear example contains a small foreground manipulation embedded in a broad outdoor scene. In each case, the heatmap remains concentrated around the altered object rather than spreading across the complete image. The predicted mask removes most low-response background pixels, and the overlay provides a direct check of whether the retained region corresponds to the object discussed in the explanation. This sequence makes two possible failure points visible to an examiner: an inaccurate heatmap would indicate insufficient pixel evidence, while a poorly selected mask would indicate a proposal or ranking error even if the initial response were informative.

Viewed together, the two panels clarify the complementary outputs of the three-agent design. Agent A contributes the image-level semantic hypothesis and the description of the suspicious content, Agent B supplies the spatial response and selected mask, and Agent C determines whether the local evidence is sufficient to retain a locally tampered decision. The qualitative figure does not replace the Agent ablation or decision-path audit; instead, it connects their numerical findings to observable intermediate outputs. In particular, the displayed heatmap, mask, and overlay allow the final category to be inspected against a concrete spatial support rather than only an image-level score.

V. Conclusion

This paper presented MADL, a multi-agent framework that organizes image forgery detection and localization as heterogeneous evidence adjudication. Agent A performs semantic verification and image-level three-class classification, Agent B extracts pixel-level traces and selects localization masks, and Agent C explicitly reconciles their outputs through agreement and conflict states. Experiments on SID-Set demonstrate competitive three-class classification, strong IoU-based localization, complementary agent roles, and traceable decision paths. Nevertheless, cross-dataset generalization requires broader evaluation, very small or fragmented manipulations remain difficult to localize, and repeated model calls increase computation and latency. Future work will explore weak-trace and fine-grained region modeling, efficient collaboration between large and compact forensic models, and reinforcement-learning-based multi-agent coordination under unseen editing and post-processing conditions.

References

[1] V. Bazyleva, N. Bonettini, and G. Bharaj, "X-edit: Detecting and localizing edits in images altered by text-guided diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2025, pp. 2720–2730.

[2] Z. Huang, J. Hu, X. Li, Y. He, X. Zhao, B. Peng, B. Wu, X. Huang, and G. Cheng, "Sida: Social media image deepfake detection, localization and explanation with large multimodal model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 28 831–28 841.

[3] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1053–1061.

[4] Y. Wu, W. Abd-Almageed, and P. Natarajan, "Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 9535–9544.

[5] S. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8692–8701.

[6] Z. Liu, X. Qi, and P. H. S. Torr, "Global texture enhancement for fake face detection in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8057–8066.

[7] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "Dire for diffusion-generated image detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 22 388–22 398.

[8] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 24 480–24 489.

[9] Y. Chang, C. Yeh, W. Chiu, and N. Yu, "Antifakeprompt: Prompt-tuned vision-language models are fake image detectors," *arXiv preprint arXiv:2310.17419*, 2023.

[10] J. Chen, X. Liao, W. Wang, Z. Qian, Z. Qin, and Y. Wang, "SNIS: A signal noise separation-based network for post-processed image forgery detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 2, pp. 935–951, 2023.

[11] J. Chen, X. Liao, Z. Qian, and Z. Qin, "PRest-Net: Multi-domain probability estimation network for robust image forgery detection," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 21, no. 3, pp. 89:1–89:20, 2025.

[12] J. Chen, X. Liao, W. Wang, and Z. Qin, "Identification of image global processing operator chain based on feature decoupling," *Information Sciences*, vol. 637, p. 118961, 2023.

[13] X. Chen, C. Dong, J. Ji, J. Cao, and X. Li, "Image manipulation detection by multi-view multi-scale supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 14 165–14 173.

[14] X. Liu, Y. Liu, J. Chen, and X. Liu, "Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7505–7517, 2022.

[15] X. Guo, X. Liu, Z. Ren, S. Grosz, I. Masi, and X. Liu, "Hierarchical fine-grained image forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 3155–3165.

[16] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, "Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 20 606–20 615.

[17] Z. Yu, J. Ni, Y. Lin, H. Deng, and B. Li, "Diffforensics: Leveraging diffusion prior to image forgery detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 12 765–12 774.

[18] Y. Li, F. Cheng, W. Yu, G. Wang, G. Luo, and Y. Zhu, "Adaifl: Adaptive image forgery localization via a dynamic and importance-aware transformer network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 477–493.

[19] H. Zhu, W. Zhu, G. Cao, and X. Huang, "Progressive feedback-enhanced transformer for image forgery localization," *arXiv preprint arXiv:2311.08910*, 2023.

[20] M. Wang, S. Niu, and J. Zhang, "Forgdiffuser: General image forgery localization with diffusion models," in *Proc. 33rd Int. Joint Conf. Artif. Intell. (IJCAI)*, 2024, pp. 1954–1962.

[21] J. Hu, X. Liao, D. Gao, S. Tsutsui, Q. Wang, Z. Qin, and M. Z. Shou, "Delocate: Detection and localization for deepfake videos with randomly-located tampered traces," in *Proc. 33rd Int. Joint Conf. Artif. Intell. (IJCAI)*, 2024.

[22] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia, "Lisa: Reasoning segmentation via large language model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 9579–9589.

[23] Z. Xu, X. Zhang, R. Li, Z. Tang, Q. Huang, and J. Zhang, "Fakeshield: Explainable image forgery detection and localization via multi-modal large language models," *arXiv preprint arXiv:2410.02761*, 2024.

[24] R. Kundu, A. Balachandran, and A. K. Roy-Chowdhury, "Truthlens: Explainable deepfake detection for face manipulated and fully synthetic data," *arXiv preprint arXiv:2503.15867*, 2025.

[25] Y. Wu, Z. Zhou, Y. Luo, J. Ji, K. Yan, and S. Ding, "Vlforgery face triad: Detection, localization and attribution via multimodal large language models," *arXiv preprint arXiv:2503.06142*, 2025.

[26] Z. Zhou, Y. Luo, Y. Wu, K. Sun, J. Ji, K. Yan, S. Ding, X. Sun, Y. Wu, and R. Ji, "Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2025, pp. 18 746–18 758.

[27] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang, "Genimage: A million-scale benchmark for detecting ai-generated image," *arXiv preprint arXiv:2306.08571*, 2023.

[28] C. Miao *et al.*, "DDL: A large-scale datasets for deepfake detection and localization in diversified real-world scenarios," *arXiv preprint arXiv:2506.23292*, 2025.

[29] B. Liu, F. Yang, B. Xiao, W. Li, G. Huang, and X. Gao, "Detecting generated images by real images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[30] P. Gao, J. Qin, X. Xiang, and Y. Tan, "Robust and privacy-preserving feature extractor for perturbed images," *Pattern Recognition*, vol. 161, p. 111202, 2025.

[31] J. Liao, J. Qin, Y. Luo, W. Pan, X. Xiang, and Y. Tan, "CLME: Robust screen-shooting watermarking with contrastive learning and mask-guided embedding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 10, pp. 9922–9935, 2025.

[32] Z. Liu, J. Qin, X. Xiang, Y. Luo, and Y. Tan, "Post-encoding enhancement: A screen-shooting resistant watermarking scheme with feature enhancement and hybrid distortion simulation," *Knowledge-Based Systems*, vol. 322, p. 113673, 2025.

[33] J. Liao, J. Qin, Y. Luo, and X. Xiang, "SSRW-INN: An invertible neural network for screen shooting robust watermarking," *Pattern Recognition*, vol. 180, p. 113929, 2026.

[34] L. Luo, X. Xiang, J. Qin, W. Pan, Y. Luo, and Y. Tan, "DCL-Net: Decoupled contrastive learning network for image manipulation localization," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2026.

[35] L. Luo, X. Xiang, J. Qin, and Y. Tan, "Dual-domain multi-scale and edge-guided network for image manipulation localization," *Expert Systems with Applications*, p. 133693, 2026.

[36] W. Xiong and V. Kuklina, "ForgeryGPT: Cross-domain face forgery detection using large vision-language models," in *Neural Information Processing*, ser. Lecture Notes in Computer Science. Singapore: Springer, 2025, vol. 15293.

[37] Y. Wang *et al.*, "ForgeryVCR: Visual-centric reasoning via efficient forensic tools in MLLMs for image forgery detection and localization," *arXiv preprint arXiv:2602.14098*, 2026.

[38] Z. Liu *et al.*, "FOCA: Frequency-oriented cross-domain forgery detection, localization and explanation via multi-modal large language model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026, pp. 12 332–12 336.

[39] S. Guo, J. Cui, and R. Hong, "Rethinking VLMs for image forgery detection and localization," *arXiv preprint arXiv:2603.12930*, 2026.

[40] G. Deng, Y. Liu, Y. Li, H. Wang, and Y. Liu, "Pentestgpt: An llm-empowered automatic penetration testing tool," in *Proc. 33rd USENIX Secur. Symp.*, 2024.

[41] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch, "Improving factuality and reasoning in language models through multiagent debate," in *Proc. 41st Int. Conf. Mach. Learn. (ICML)*, 2024.

[42] M. Ghafourian, H. Aboutorab, and D. Giannoula, "Unishield: A unified framework for multi-domain image forgery detection," *arXiv preprint arXiv:2412.01785*, 2024.

[43] P. Gao, J. Qin, X. Xiang, and Y. Tan, "Knowledge distillation from relative distribution," *Expert Systems with Applications*, vol. 284, p. 127736, 2025.

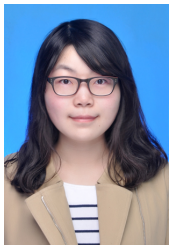
[44] B. Zhang, J. Qin, X. Xiang, and Y. Tan, "Learning continuation: Integrating past knowledge for contrastive distillation," *Knowledge-Based Systems*, vol. 304, p. 112573, 2024.

[45] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 4015–4026.
[46] J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 868–882, 2012.
[47] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 11 966–11 976.
[48] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-vl technical report," *arXiv preprint arXiv:2502.13923*, 2025.
[49] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

Dashan Qing received the M.S. degree from the School of Computer Science and Technology, Central South University of Forestry and Technology, Changsha, China, in 2024. He is currently pursuing the Ph.D. degree with the College of Cyber Science and Technology, Hunan University, Changsha, China. His research interests include digital image forensics and multimedia security.

Xin Liao (Senior Member, IEEE) received the B.E. and Ph.D. degrees in information security from the Beijing University of Posts and Telecommunications, Beijing, China, in 2007 and 2012, respectively. He is currently a Professor with the College of Cyber Science and Technology, Hunan University, Changsha, China, where he also serves as the Vice Dean. He was a Postdoctoral Fellow with the Institute of Software, Chinese Academy of Sciences, Beijing, China, and a Research Associate with The University of Hong Kong, Hong Kong. From 2016 to 2017, he was a Visiting Scholar with the University of Maryland, College Park, MD, USA. His research interests include multimedia forensics, artificial intelligence security, and watermarking. He is the Vice Chair of the Technical Committee on Multimedia Security and Forensics of the Asia-Pacific Signal and Information Processing Association, a member of the Technical Committee on Computer Forensics of the Chinese Institute of Electronics, and a member of the Technical Committee on Digital Forensics and Security of the China Society of Image and Graphics. He serves as an Associate Editor for the IEEE Signal Processing Magazine.

Chengyu Wang received the M.S. degree in computer science and technology and the Ph.D. degree in cybersecurity from the National University of Defense Technology, Changsha, China, in 2021 and 2025, respectively. He is currently an Assistant Professor with the College of Cyber Science and Technology, Hunan University, Changsha, China. His research interests include time-series analysis and artificial intelligence security.



Jiaxin Chen received the B.S. degree from Central China Normal University, Wuhan, China, in 2017, and the Ph.D. degree from Hunan University, Changsha, China, in 2023. She is currently a Lecturer with the School of Computer Science and Technology, Changsha University of Science and Technology, Changsha, China. Her research interests include multimedia forensics and deepfake detection. She is a member of the Technical Committee on Media Security and Forensics of the Asia-Pacific Signal and Information Processing Association and the Technical Committee on Digital Forensics and Security of the China Society of Image and Graphics.