RoGRE: Image Manipulation Localization as Controlled Hypothesis Revision

Anonymous authors
Paper under double-blind review

Abstract

Image manipulation localization (IML) requires combining heterogeneous evidence with different localization capabilities and failure modes. Existing methods typically fuse such representations as equal-status features, obscuring how each source should modify the current prediction. We instead formulate IML as *controlled hypothesis revision*: localization begins from an initial estimate, and complementary evidence progressively corrects its remaining errors. Based on this view, we introduce RoGRE, which maintains a shared logit-valued localization state and revises it through role-specific stages. DINOv3 first establishes an initial hypothesis, SAM2 features correct its spatial organization, and forensic residual features refine manipulation evidence that remains weakly represented. Each refinement proposes a bounded signed correction whose magnitude is regulated by a gate conditioned on both the incoming state and stage-specific evidence, allowing informative revisions while limiting unreliable updates. The evolved state is finally decoded by SAM2 into the manipulation mask. RoGRE achieves mean pixel-level F1 scores of 0.888 and 0.813 under the CAT-Net and MVSS-Net protocols, respectively. Ablations further validate the correction order and state-dependent gating, supporting controlled hypothesis revision as an effective approach to heterogeneous evidence integration for IML.

1 Introduction

Image manipulation localization (IML) identifies manipulated regions at the pixel level across operations such as splicing, copy-move, inpainting, and post-processing (Farid, 2009; Wu et al., 2018). Unlike semantic segmentation, manipulation is not defined by object category or appearance: an edited region may remain semantically plausible and spatially coherent while carrying subtle traces of image formation or editing. Localization therefore draws on heterogeneous evidence, from visual context and spatial structure to low-level forensic statistics.

A long line of work has explored complementary information for IML. Traditional forensic methods exploit sensor noise, constrained prediction errors, resampling artifacts, and high-frequency residuals (Lukáš et al., 2006; Bayar & Stamm, 2016; Fridrich & Kodovský, 2012). Learning-based methods incorporate semantic context, compression cues, boundaries, noise patterns, and cross-region consistency (Wu et al., 2019; Chen et al., 2021; Kwon et al., 2022; Guillaro et al., 2023). Foundation models further provide transferable dense representations and segmentation priors (Siméoni et al., 2025; Ravi et al., 2025; Lek et al., 2026; Zhang et al., 2025; Yu et al., 2026a). These developments broaden the evidence available to IML, but also raise the question of how different sources should influence a localization estimate.

The sources differ in both their strengths and failure modes. High-level visual representations capture global context but may miss subtle manipulation traces or fine boundaries. Segmentation-oriented features provide spatial organization but favor object structure over forensic inconsistency. Forensic residuals reveal weak editing traces but can also respond to benign texture, compression, or acquisition artifacts. When these sources are treated as equal-status features, a shared fusion space must reconcile these differences. The challenge therefore lies not only in what each source contributes, but also in what it should correct and how strongly it should affect the current prediction.

We view this integration problem as *controlled hypothesis revision*. Once a meaningful localization estimate exists, subsequent evidence need not solve localization again; instead, it can revise the

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.

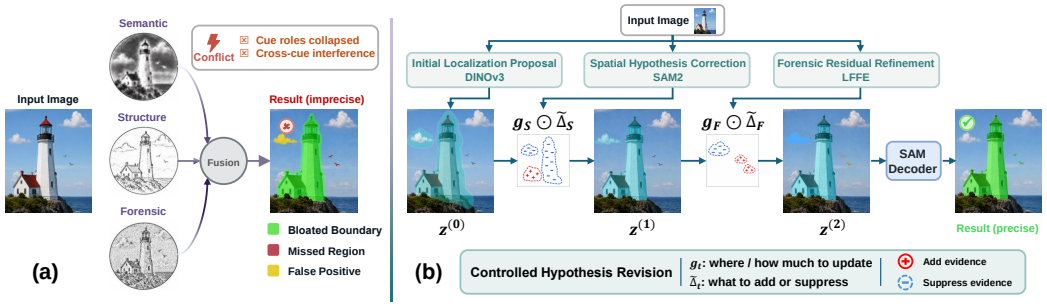


Figure 1: **From (a) Equal-status Fusion to (b) Controlled Hypothesis Revision.** RoGRE integrates heterogeneous evidence by revising a shared localization hypothesis. DINOv3 establishes the initial hypothesis, SAM2 features correct its spatial organization, and forensic features refine residual manipulation traces. Gated signed updates regulate each correction before final mask decoding.

current hypothesis according to complementary roles. Structural evidence can address spatial organization and boundaries, while forensic evidence can target manipulation traces that remain weakly represented. The current hypothesis then serves as a reference to regulate each correction. Under this view, the central question shifts from "*how should heterogeneous features be fused?*" to "*what should each source of evidence be allowed to correct, and how strongly should it be allowed to change the current hypothesis?*".

Based on this formulation, we introduce **RoGRE**, a **Ro**le-structured **G**ated **R**esidual **E**volution framework that implements this controlled-revision perspective through a shared logit-valued localization state (Fig. 1(b)). RoGRE comprises three stages: ***Initial Localization Proposal***, ***Spatial Hypothesis Correction***, and ***Forensic Residual Refinement***. DINOv3 first establishes a coarse but meaningful localization hypothesis from transferable visual representations. Conditioned on this reference, SAM2 image features correct its spatial organization and boundary structure using structural and high-resolution information, rather than independently predicting a new mask. Finally, a lightweight forensic fingerprint encoder supplies low-level evidence to refine residual manipulation traces that remain insufficiently captured, with gated residual updates preventing noisy forensic cues from indiscriminately overwriting the structured hypothesis.

At each correction stage, the incoming evidence proposes a bounded signed residual specifying *what* should change. A learned gate, conditioned on both the current state and the incoming evidence, determines *how much* of that proposal is accepted. Positive and negative residuals can strengthen or suppress manipulation evidence, and the gate regulates their contribution to the shared state. This makes each correction explicitly state-dependent. The evolved state is then passed as a dense prompt to the SAM2 decoder to produce the final manipulation mask.

Under the CAT-Net and MVSS-Net protocols, RoGRE achieves mean pixel-level F1 scores of 0.888 and 0.813, respectively, outperforming the reported competing methods. Component, ordering, and gating ablations further show lower performance when updates are concatenated, applied in parallel, reversed, or made independent of the current state. These results support the effectiveness of RoGRE's ordered, state-dependent evidence integration.

Our contributions are:

- We formulate heterogeneous evidence integration for IML as *controlled hypothesis revision*, assigning complementary error-correction roles to visual, structural, and forensic representations.
- We introduce RoGRE, which maintains a shared localization state through initial proposal, spatial correction, and forensic refinement, organizing the three evidence sources into an ordered localization process.
- We develop a gated residual mechanism that separates the content of a correction from its accepted magnitude, using bounded signed residuals and gates conditioned on the current state and incoming evidence.
- Experiments under two evaluation protocols, together with component, ordering, and gating ablations, demonstrate the effectiveness of RoGRE and support the proposed evidence integration design.

2 Related Work

Image Manipulation Localization. IML has progressed from handcrafted resampling and camera-trace analysis (Popescu & Farid, 2005; Lukáš et al., 2006) to learned representations for prediction errors, anomalous traces, camera fingerprints, and source–target consistency (Zhou et al., 2018; Bayar & Stamm, 2016; Wu et al., 2019; Cozzolino & Verdoliva, 2020; Wu et al., 2018; Mazaheri et al., 2019). Recent systems combine compression, boundaries, correlations, object context, and uncertainty (Chen et al., 2021; Kwon et al., 2022; Liu et al., 2022; Guillaro et al., 2023; Wang et al., 2022; Ji et al., 2023). TransForensics refines outputs from different branches (Hao et al., 2021), PIM-Net progressively fuses multi-scale inconsistency features (Bai et al., 2025), and Vision Mamba offers another route to multi-scale aggregation (Guo et al., 2025). These approaches motivate the use of complementary cues, but combining representations does not by itself specify the localization role of each cue. RoGRE assigns Initial Localization Proposal, Spatial Hypothesis Correction, and Forensic Residual Refinement to successive updates of one localization state.

Foundation Models for Dense Prediction and Image Forensics. DINO and DINOv3 provide transferable visual representations (Caron et al., 2021; Siméoni et al., 2025), while transformer decoders support flexible mask prediction (Strudel et al., 2021; Cheng et al., 2022). SAM and SAM2 expose prompt-conditioned segmentation (Kirillov et al., 2023; Ravi et al., 2025), and long-range interaction, efficient promptable segmentation, and parameter-efficient adaptation provide complementary architectural tools (Huang et al., 2026; Xiong et al., 2024; Hu et al., 2022). DINOv3-IML maps adapted DINOv3 tokens directly to a mask with a lightweight convolutional head (Yu et al., 2026a). IMDPrompter generates and reconciles SAM prompts from RGB and forensic views (Zhang et al., 2025), whereas Detective SAM derives SAM2 prompts from perturbation-based forensic clues (Lek et al., 2026). In RoGRE, DINOv3 and SAM2 instead take different positions along a single localization trajectory: DINOv3 forms the Initial Localization Proposal, SAM2 features perform Spatial Hypothesis Correction, forensic features perform Forensic Residual Refinement, and the resulting state prompts the SAM2 mask decoder.

3 Method

3.1 Overview

Given an RGB image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$, our goal is to predict a binary manipulation mask $\mathbf{y} \in \{0,1\}^{H\times W}$. RoGRE organizes heterogeneous cues by their localization roles instead of merging them into one feature tensor or producing a mask from each branch. As illustrated in Figure 2, DINOv3 forms the ***Initial Localization Proposal***, SAM2 image features perform ***Spatial Hypothesis Correction***, and the lightweight forensic fingerprint encoder (LFFE) performs ***Forensic Residual Refinement***. The latter two stages apply gated corrections to the same logit-valued localization state:

$$\mathbf{z}^{(1)} = \mathbf{z}^{(0)} + \Delta_{\mathrm{S}}, \qquad \mathbf{z}^{(2)} = \mathbf{z}^{(1)} + \Delta_{\mathrm{F}}, \tag{1}$$

where $\mathbf{z}^{(0)}$ denotes the DINOv3 semantic logit map, Δ_{S} and Δ_{F} are the signed corrections produced by the SAM2 and forensic stages, respectively, and $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$ denote the localization states after the first and second corrections. The state after two gated residual updates, $\mathbf{z}^{(2)}$, is subsequently passed to SAM2 as a dense mask prompt.

This sequence realizes controlled hypothesis revision through role-specific updates. The spatial stage turns $\mathbf{z}^{(0)}$ into $\mathbf{z}^{(1)}$; the forensic stage starts from that spatially organized hypothesis and produces $\mathbf{z}^{(2)}$. Bounded, state-conditioned updates preserve the shared trajectory while allowing either correction to strengthen or weaken local manipulation evidence. These intermediate states are not independent mask predictions; the final SAM2 decoder combines $\mathbf{z}^{(2)}$ with image features to produce the output mask.

3.2 Role-Structured State Evolution

Initial Localization Proposal. We adapt a pretrained DINOv3 encoder (Siméoni et al., 2025) with low-rank adapters (LoRA) (Hu et al., 2022) to produce the initial semantic localization state.

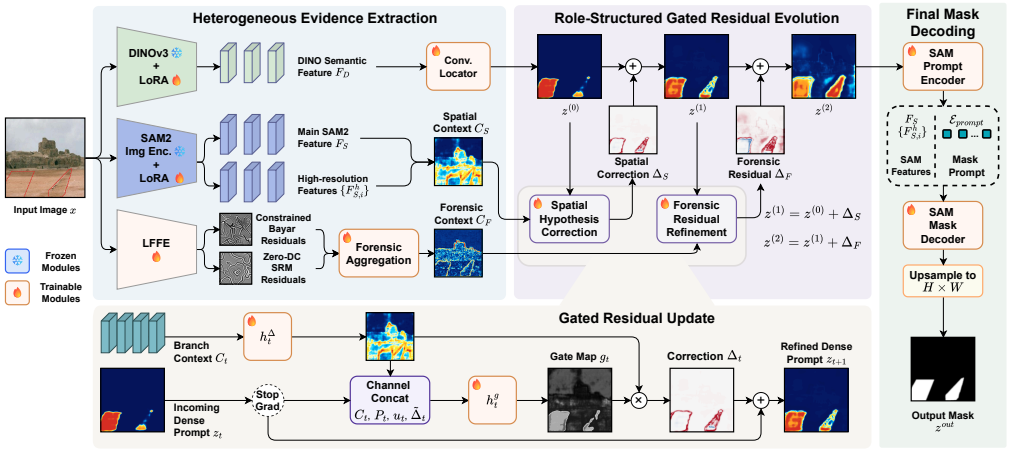


Figure 2: **RoGRE Architecture.** DINOv3 forms the Initial Localization Proposal, SAM2 image features perform Spatial Hypothesis Correction, and forensic fingerprints perform Forensic Residual Refinement. Each of the latter two stages applies a bounded, state-conditioned update before the evolved state is passed to SAM2 for final mask decoding.

A lightweight convolutional locator maps its final feature map $\mathbf{F}_{\mathrm{D}}$ to a logit map, which is resized to the resolution expected by the SAM2 dense prompt:

$$\mathbf{z}^{(0)} = \mathcal{U}_p(h_{\mathrm{D}}(\mathbf{F}_{\mathrm{D}}))\,. \tag{2}$$

Here, h_{D} is the learnable convolutional locator and $\mathcal{U}_p$ denotes bilinear interpolation to dense-prompt resolution p. Thus, $\mathbf{z}^{(0)}$ is a logit map at the same resolution as the subsequent localization states, rather than an intermediate feature tensor.

This stage establishes a global localization hypothesis rather than resolving every fine-scale manipulation detail. The probability and uncertainty derived from $\mathbf{z}^{(0)}$ provide a reference for the next gated correction. We supervise $\mathbf{z}^{(0)}$ with the mask loss in Section 3.5, so subsequent stages update a meaningful estimate rather than an unconstrained latent prompt.

Spatial Hypothesis Correction. The first correction stage exploits the main and high-resolution image features $\mathbf{F}_{\mathrm{S}}$ and $\mathbf{F}^h_{\mathrm{S},i}$ of the SAM2 (Ravi et al., 2025) image encoder. After channel projection and spatial resizing, these features are combined to form the spatial context $\mathbf{C}_{\mathrm{S}}$. It takes $\mathbf{z}^{(0)}$ as the incoming hypothesis and produces the spatially corrected state $\mathbf{z}^{(1)}$ through the shared gated residual mechanism defined in Section 3.3.

Rather than acting as an independent SAM2 segmentation branch, this stage corrects the spatial organization of the current DINOv3 proposal. The main SAM2 embedding provides object-level structural context, whereas its high-resolution features preserve finer spatial details. Their complementary information aligns the proposal with object continuity, local structure, and manipulation boundaries.

Forensic Residual Refinement. The second stage extracts low-level forensic fingerprints using our lightweight forensic fingerprint encoder (LFFE). LFFE combines constrained Bayar prediction errors (Bayar & Stamm, 2016) with learnable zero-DC SRM residuals (Fridrich & Kodovský, 2012), followed by shallow local–context forensic feature aggregation. Importantly, LFFE contains no raw-RGB bypass, encouraging its forensic context $\mathbf{C}_{\mathrm{F}}$ to preserve processing inconsistencies rather than semantic appearance.

The forensic stage does not predict a second mask from $\mathbf{C}_{\mathrm{F}}$ alone. It takes the spatially corrected state $\mathbf{z}^{(1)}$ as its reference and produces the final evolved state $\mathbf{z}^{(2)}$ through the same gated residual mechanism. Neutral initialization makes the initial forensic update zero, preserving the semantic–spatial pathway. During training, the stage learns to correct remaining errors using residual evidence. Conditioning the update on $\mathbf{z}^{(1)}$ and bounding its magnitude are especially relevant here: acquisi-

tion noise and benign post-processing can also produce strong residual responses. Detailed filter construction, channel dimensions, and initialization settings are provided in Appendix A.

3.3 Gated Residual Update Mechanism

The Spatial Hypothesis Correction and Forensic Residual Refinement stages share one update mechanism while using role-specific context features. Let $t \in \{\mathrm{S}, \mathrm{F}\}$ index the spatial and forensic stages, respectively. At either stage, $\mathbf{z}_t$ is the incoming logit map and $\mathbf{C}_t$ is the branch-specific context feature. All quantities below are aligned to the dense-prompt resolution. The stage first predicts a bounded signed candidate, summarizes the current state as a probability $\mathbf{p}_t$ and an uncertainty map $\mathbf{u}_t$, and then uses these inputs to determine the applied correction:

$$\begin{aligned}
&\widetilde{\Delta}_t = \alpha_t \tanh\big(h_t^{\Delta}(\mathbf{C}_t)\big)\,, \\
&\mathbf{p}_t = \sigma(\mathrm{sg}(\mathbf{z}_t))\,, \qquad \mathbf{u}_t = \frac{-\mathbf{p}_t \log(\mathbf{p}_t + \epsilon) - (1-\mathbf{p}_t)\log(1-\mathbf{p}_t+\epsilon)}{\log 2}, \\
&\mathbf{g}_t = \sigma\Big(h_t^{g}\Big(\mathrm{Concat}[\mathbf{C}_t, \mathbf{p}_t, \mathbf{u}_t, \widetilde{\Delta}_t]\Big)\Big)\,, \qquad \Delta_t = \mathbf{g}_t \odot \widetilde{\Delta}_t.
\end{aligned} \tag{3}$$

Here, $\mathbf{z}_{\mathrm{S}} = \mathbf{z}^{(0)}$ and $\mathbf{z}_{\mathrm{F}} = \mathbf{z}^{(1)}$; consequently, the corresponding applied corrections are Δ_{S} and Δ_{F} in equation 1. The predictor h_t^{Δ} maps $\mathbf{C}_t$ to the correction map. Since $\alpha_t > 0$ and $\tanh(\cdot) \in [-1, 1]$, each element of $\widetilde{\Delta}_t$ lies in $[-\alpha_t, \alpha_t]$: positive values increase the manipulation logit, whereas negative values decrease it.

For state conditioning, $\mathrm{sg}(\cdot)$ denotes stop-gradient and $\sigma(\cdot)$ denotes the sigmoid function. Thus, $\mathbf{p}_t$ is the detached foreground-probability map derived from $\mathbf{z}_t$, and $\mathbf{u}_t$ is its normalized binary-entropy map. The constant $\epsilon > 0$ ensures numerical stability, and the division by $\log 2$ places $\mathbf{u}_t$ in $[0, 1]$, with larger values indicating a less decisive incoming state. Here, $\mathrm{Concat}[\cdot]$ denotes channel-wise concatenation. The gating network h_t^g therefore receives the branch context, current foreground probability, state uncertainty, and proposed signed change, and produces the one-channel gate $\mathbf{g}_t \in [0, 1]$. Element-wise multiplication $\odot$ scales the candidate to form Δ_t: the candidate specifies a signed change, while the gate modulates its magnitude using branch evidence and the incoming state. This parameterization bounds each prompt update and makes its magnitude depend on the current localization hypothesis. Only the conditioning path is detached. The additive state path in equation 1 remains differentiable, so each stage can still learn how its correction improves the final localization state.

3.4 Final Mask Decoding

After Spatial Hypothesis Correction and Forensic Residual Refinement, the evolved state $\mathbf{z}^{(2)}$ is passed to the SAM2 prompt encoder as a dense mask prompt:

$$\mathcal{E}_{\mathrm{prompt}} = E_{\mathrm{P}}(\mathbf{z}^{(2)}), \qquad \mathbf{z}^{\mathrm{out}} = \mathcal{U}_{H,W}\big[D_{\mathrm{M}}\big(\mathbf{F}_{\mathrm{S}}, \mathcal{E}_{\mathrm{prompt}}, \{\mathbf{F}_{\mathrm{S},i}^{h}\}\big)\big]\,. \tag{4}$$

Here, E_{P} denotes the SAM2 prompt encoder, while $\mathcal{E}_{\mathrm{prompt}}$ denotes the prompt embedding that encodes the evolved localization state $\mathbf{z}^{(2)}$. D_{M} denotes the SAM2 mask decoder, $\mathbf{F}_{\mathrm{S}}$ is the main SAM2 image embedding, and $\mathbf{F}_{\mathrm{S},i}^{h}$ denotes its i-th high-resolution feature map. Finally, $\mathcal{U}_{H,W}$ resizes the decoder prediction to the original image resolution, producing the final localization logits $\mathbf{z}^{\mathrm{out}}$. The decoder converts the state produced by the three role-specific stages into a mask using SAM2 image features. Appendix D examines how this final step affects false positives and spatial coherence.

3.5 Role-Aligned Training Objectives

The objectives follow the same role order used at inference: Initial Localization Proposal, boundary-aware Spatial Hypothesis Correction, Forensic Residual Refinement, and final mask prediction. For both the final prediction and the initial DINOv3 proposal, we define

$$\mathcal{L}_{\mathrm{mask}}(\mathbf{z}, \mathbf{y}) = 0.2\,\mathcal{L}_{\mathrm{wbce}}(\mathbf{z}, \mathbf{y}) + 0.8\,\mathcal{L}_{\mathrm{dice}}(\mathbf{z}, \mathbf{y}), \tag{5}$$

where $\mathbf{z}$ denotes the supervised logit map, $\mathbf{y}$ is the resolution-matched binary target, $\mathcal{L}_{\mathrm{wbce}}$ denotes weighted binary cross-entropy, and $\mathcal{L}_{\mathrm{dice}}$ denotes the squared Dice loss (Milletari et al., 2016). The

Dice term is evaluated only for samples containing manipulated foreground. The detailed definitions of $\mathcal{L}_{\mathrm{wbce}}$ and $\mathcal{L}_{\mathrm{dice}}$ are provided in Appendix B.

Beyond mask supervision, the intermediate spatial state $\mathbf{z}^{(1)}$ receives boundary-focused supervision through $\mathcal{L}_{\mathrm{bdry}}$, encouraging the SAM2 correction to improve geometry around manipulation contours. The forensic auxiliary loss $\mathcal{L}_{\mathrm{for}}$ is evaluated over complementary soft residual-error supports derived from the detached probability of $\mathbf{z}^{(1)}$. These supports emphasize missed manipulated pixels and false-positive responses in the incoming state, aligning Forensic Residual Refinement with remaining localization errors rather than an independent mask prediction.

With the weights used in the final model, the complete objective is

$$\boxed{\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{mask}}(\mathbf{z}^{\mathrm{out}}, \mathbf{y}) + \lambda_D \, \mathcal{L}_{\mathrm{mask}}(\mathbf{z}^{(0)}, \mathbf{y}) + \lambda_S \, \mathcal{L}_{\mathrm{bdry}}(\mathbf{z}^{(1)}, \mathbf{y}) + \lambda_F \, \mathcal{L}_{\mathrm{for}}.} \tag{6}$$

The first term supervises the final decoded localization logits $\mathbf{z}^{\mathrm{out}}$, while the second maintains a meaningful Initial Localization Proposal $\mathbf{z}^{(0)}$. $\mathcal{L}_{\mathrm{bdry}}$ supervises the spatially corrected state $\mathbf{z}^{(1)}$ around manipulation boundaries, and $\mathcal{L}_{\mathrm{for}}$ supervises the signed update from Forensic Residual Refinement over the remaining residual-error supports. The final mask objective has unit weight, while each auxiliary objective is assigned a weight of 0.1.

The full loss definitions and implementation details are provided in Appendix B. Overall, the objective supports each position along the shared localization trajectory: Initial Localization Proposal, Spatial Hypothesis Correction, Forensic Residual Refinement, and the final decoded mask.

4 Experiments

4.1 Experimental Setup

Protocols and Metrics. We evaluate pixel-level IML under the four-source CAT-Net protocol used by DINOv3-IML (Kwon et al., 2022; Yu et al., 2026a) and the MVSS-Net protocol (Chen et al., 2021). Our CAT-Net setting trains on CASIA-v2 (Dong et al., 2013), FantasticReality (Kniaz et al., 2019), IMD2020 (Novozamsky et al., 2020), and TampCOCO (Kwon et al., 2022); it omits CompRAISE from the full CAT-Net training mixture. We then evaluate on CASIAv1 (Dong et al., 2013), Columbia (Hsu & Chang, 2006), NIST16 (Fiscus et al., 2017), and Coverage (Wen et al., 2016). The MVSS-Net protocol trains on CASIA-v2 and evaluates on CASIAv1, Columbia, NIST16, Coverage, and IMD2020. We apply the same threshold of 0.5 to every test set and report pixel-level F1 on tampered images for each dataset.

Implementation. RoGRE combines DINOv3 ViT-L/16 (Siméoni et al., 2025) and SAM2 Hiera-B+ (Ravi et al., 2025) at 512×512 resolution. We insert rank-32 LoRA adapters with scaling factor 64 into the query, key, and value projections of both image encoders (Hu et al., 2022); the pretrained encoder weights remain frozen apart from these adapters, while the SAM2 prompt encoder and mask decoder are trained. LFFE extracts forensic fingerprints from Bayar-constrained prediction errors and learnable SRM residual banks (Bayar & Stamm, 2016; Fridrich & Kodovský, 2012). Appendix A provides the complete architecture and optimization settings. We repeat our experiments five times with different random seeds and report the mean F1.

4.2 Main Results

Table 1 reports results under the two protocols separately.[1] Under the CAT-Net protocol, RoGRE achieves a mean pixel-level F1 of 0.888, compared with 0.847 for the best competing mean. It performs best on CASIAv1 (0.942), NIST16 (0.759), and Coverage (0.917). Columbia remains more challenging, where RoGRE obtains 0.935, below APSC-Net (0.966) and DINOv3-IML (0.941). The overall mean thus reflects broad improvements without implying consistent superiority on every dataset. Under the narrower MVSS-Net protocol, RoGRE achieves the best F1 on all five test datasets, with a mean of 0.813 compared with 0.774 for the best competing mean. The gains across

[1] IMDPrompter (Zhang et al., 2025) has not released their code and reports only under the MVSS-Net protocol, while Detective SAM (Lek et al., 2026) provides inference code only and targets AIGC tampering under a different setting. Thus, neither supports reproducible comparison under both protocols.

Table 1: **Cross-Protocol Manipulation Localization.** Pixel-level F1 on tampered images under the four-source CAT-Net (top) and MVSS-Net (bottom) protocols. Best results in each protocol are shown in **bold**, and second-best results are underlined.

Method	Venue	Coverage	Columbia	NIST16	CASIAv1	IMD2020	Avg.
Four-Source CAT-Net Protocol Evaluation							
MVSS-Net (Chen et al., 2021)	ICCV'21	0.482	0.740	0.336	0.583	–	0.535
CAT-Net (Kwon et al., 2022)	IJCV'22	0.427	0.915	0.252	0.808	–	0.601
PSCC-Net (Liu et al., 2022)	TCSVT'22	0.448	0.884	0.346	0.630	–	0.577
TruFor (Guillaro et al., 2023)	CVPR'23	0.457	0.885	0.348	0.818	–	0.627
IML-ViT (Ma et al., 2023)	arXiv'23	0.654	0.948	0.501	0.795	–	0.725
APSC-Net (Qu et al., 2024)	CVPR'24	0.523	**0.966**	0.436	0.837	–	0.691
SparseViT (Su et al., 2025)	AAAI'25	0.513	0.959	0.384	0.827	–	0.671
Mesorch (Zhu et al., 2025)	AAAI'25	0.586	0.890	0.392	0.840	–	0.677
Co-Transformers (Zhang et al., 2026)	AAAI'26	0.654	0.956	0.455	0.855	–	0.730
EARG-Net (Yu et al., 2026b)	AAAI'26	0.755	0.924	0.589	0.498	–	0.692
RITA (Zhu et al., 2026)	CVPRF'26	0.566	0.921	0.428	0.770	–	0.671
Relay-ViT (Huang et al., 2026)	ICLR'26	0.647	0.878	0.476	0.806	–	0.702
Relay-Seg (Huang et al., 2026)	ICLR'26	0.704	0.883	0.430	0.802	–	0.705
DINOv3-IML (Yu et al., 2026a)	arXiv'26	0.905	0.941	0.636	0.907	–	0.847
RoGRE (ours)	—	**0.917**	0.935	**0.759**	**0.942**	–	**0.888**
MVSS-Net Protocol Evaluation							
Mantra-Net (Wu et al., 2019)	CVPR'19	0.090	0.243	0.104	0.125	0.055	0.123
MVSS-Net (Chen et al., 2021)	ICCV'21	0.259	0.386	0.246	0.534	0.279	0.341
CAT-Net (Kwon et al., 2022)	IJCV'22	0.296	0.584	0.269	0.581	0.273	0.401
ObjectFormer (Wang et al., 2022)	CVPR'22	0.294	0.336	0.173	0.429	0.172	0.281
PSCC-Net (Liu et al., 2022)	TCSVT'22	0.231	0.604	0.214	0.378	0.235	0.332
NCL-IML (Zhou et al., 2023)	ICCV'23	0.225	0.446	0.260	0.502	0.237	0.334
TruFor (Guillaro et al., 2023)	CVPR'23	0.419	0.865	0.324	0.721	0.322	0.530
IML-ViT (Ma et al., 2023)	arXiv'23	0.435	0.780	0.331	0.721	0.327	0.519
SparseViT (Su et al., 2025)	AAAI'25	0.287	0.781	0.245	0.646	0.230	0.438
MaskCLIP (Wang et al., 2025)	CVPR'25	0.451	0.848	0.342	0.725	0.319	0.537
Mesorch (Zhu et al., 2025)	AAAI'25	0.326	0.726	0.343	0.740	0.269	0.481
IMDPrompter (Zhang et al., 2025)	ICLR'25	0.636	0.873	0.411	0.763	0.306	0.598
EARG-Net (Yu et al., 2026b)	AAAI'26	0.403	0.806	0.269	0.563	0.242	0.457
RITA (Zhu et al., 2026)	CVPRF'26	0.443	0.756	0.346	0.522	0.379	0.489
Relay-ViT (Huang et al., 2026)	ICLR'26	0.551	0.762	0.335	0.740	0.381	0.554
Relay-Seg (Huang et al., 2026)	ICLR'26	0.569	0.756	0.273	0.760	0.357	0.543
DINOv3-IML (Yu et al., 2026a)	arXiv'26	0.820	0.915	0.621	0.873	0.641	0.774
RoGRE (ours)	—	**0.895**	**0.934**	**0.688**	**0.886**	**0.664**	**0.813**

two training distributions are consistent with the value of keeping complementary cues in distinct roles along a shared localization trajectory.

RoGRE's design offers a consistent explanation for these results, which the ablations below test directly. DINOv3 forms the Initial Localization Proposal, SAM2 features perform Spatial Hypothesis Correction, and LFFE performs Forensic Residual Refinement. The latter two stages update the same hypothesis through bounded, state-aware gates rather than fusing all features at once. Final decoding then turns the evolved state into a mask. The component, order, and gate ablations separate the contributions of these choices.

4.3 Qualitative and Degradation Analysis

Figure 3 compares final masks at the common threshold of 0.5. The examples illustrate the spatial detail of RoGRE's final localization after Initial Localization Proposal, Spatial Hypothesis Correction, and Forensic Residual Refinement; they complement the aggregate results in Table 1.

Figure 4 shows that pixel-level F1 decreases with increasing perturbation strength on CASIAv1, while degradation remains limited across the tested JPEG compression levels. Severe blur causes the largest drop, consistent with spatial smoothing weakening the high-frequency traces available to the forensic stage. Thus, the tested compression and noise conditions affect RoGRE less than strong blur; this figure does not isolate which component accounts for that difference.

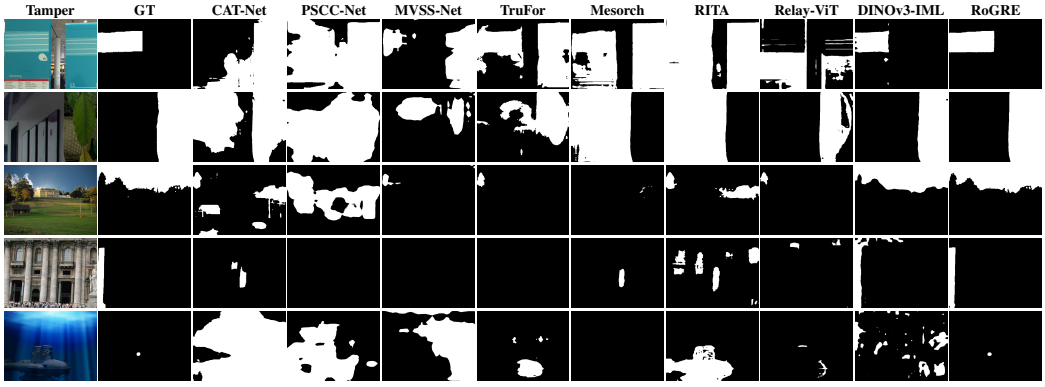


Figure 3: **Qualitative Manipulation Localization.** Final masks are compared at the common threshold of 0.5 across five datasets under the MVSS-Net protocol. Each row shows a tampered image, its ground-truth mask, baseline predictions, and the RoGRE prediction.

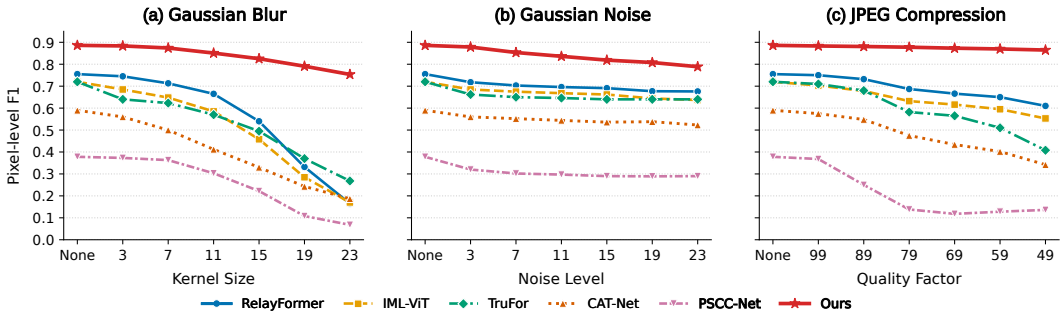


Figure 4: **Robustness to Image Degradations.** Pixel-level F1 on CASIAv1 under Gaussian blur, Gaussian noise, and JPEG compression in the MVSS-Net protocol. RoGRE is compared with methods evaluated under the same degradation settings.

4.4 Ablation Study

Table 2 tests the proposed division of labor. The DINOv3 locator (A0) supplies the Initial Localization Proposal and reaches a mean F1 of 0.827. Decoding that proposal with SAM2 (A1) raises mean F1 to 0.854, before either correction stage is added. Adding Spatial Hypothesis Correction to A1 (A2) increases it to 0.870, with gains of 1.8 points on CASIAv1 and 3.2 points on Coverage. Adding Forensic Residual Refinement to A1 instead (A3) reaches 0.872 and exceeds A2 on Columbia and NIST16, indicating that its low-level cues contribute information beyond spatial features. Applying both stages in sequence yields 0.888: 6.1 points above A0 and 3.4 points above A1.

Both residual banks contribute to the forensic encoder. Removing SRM filters (A4) lowers mean F1 by 1.1 points, while removing the Bayar-constrained branch (A5) lowers it by 0.8 points. SRM responses therefore provide the larger aggregate gain, and the Bayar branch supplies additional prediction-error evidence for the final prompt.

The gated residual evolution ablations test whether this role order and update rule matter. Concatenated updates (P0), parallel updates (P1), and reversing the forensic and spatial stages (P2) reduce mean F1 to 0.861, 0.875, and 0.868, respectively. The full Spatial Hypothesis Correction-to-Forensic Residual Refinement sequence performs best among these variants, consistent with the forensic stage benefiting from an already spatially organized hypothesis. Removing probability and uncertainty conditioning from the gates (P3) reduces mean F1 by 1.6 points; removing the gates (P4) causes a 2.8-point drop. Together, these ablations support controlled hypothesis revision through role order and state-dependent updates. Appendix D examines the prompt-to-mask transition after these updates.

Figure 5 compares the localization states with the same SAM2 decoder. Mean F1 increases from 0.867 for the Initial Localization Proposal to 0.878 after Spatial Hypothesis Correction and 0.888

Table 2: **Role and Gated Residual Evolution Ablations.** The upper block evaluates Spatial Hypothesis Correction and Forensic Residual Refinement; the lower block tests update order and state-conditioned gating under the CAT-Net protocol. F1, IoU, and accuracy are unweighted means over CASIAv1, Columbia, NIST16, and Coverage. Parenthesized values show the absolute change from RoGRE (Full) in percentage points: blue for a decrease and red for an increase.

Variant	Pixel-level F1 ↑				Four-dataset mean ↑		
	CASIAv1	Columbia	NIST16	Coverage	F1	IoU	Acc.
Component Ablations							
A0: DINOv3-only	0.897 (-4.5)	0.901 (-3.4)	0.623 (-13.6)	0.885 (-3.2)	0.827 (-6.1)	0.754 (-5.3)	0.955 (-2.1)
A1: +SAM2 decoder	0.927 (-1.5)	0.924 (-1.1)	0.681 (-7.8)	0.884 (-3.3)	0.854 (-3.4)	0.744 (-6.3)	0.951 (-2.5)
A2: +Spatial correction	0.945 (+0.3)	0.885 (-5.0)	0.734 (-2.5)	0.916 (-0.1)	0.870 (-1.8)	0.779 (-2.8)	0.968 (-0.8)
A3: +LFFE	0.935 (-0.7)	0.909 (-2.6)	0.736 (-2.3)	0.910 (-0.7)	0.872 (-1.6)	0.782 (-2.5)	0.971 (-0.5)
A4: w/o SRM	0.943 (+0.1)	0.914 (-2.1)	0.738 (-2.1)	0.911 (-0.6)	0.877 (-1.1)	0.789 (-1.8)	0.973 (-0.3)
A5: w/o Bayar	0.944 (+0.2)	0.924 (-1.1)	0.729 (-3.0)	0.923 (+0.6)	0.880 (-0.8)	0.796 (-1.1)	0.974 (-0.2)
Gated Residual Evolution Ablations							
P0: Concatenated updates	0.924 (-1.8)	0.911 (-2.4)	0.701 (-5.8)	0.906 (-1.1)	0.861 (-2.7)	0.755 (-5.2)	0.952 (-2.4)
P1: Parallel updates	0.936 (-0.6)	0.920 (-1.5)	0.730 (-2.9)	0.905 (-1.2)	0.875 (-1.3)	0.787 (-2.0)	0.972 (-0.4)
P2: F→S order	0.930 (-1.2)	0.913 (-2.2)	0.710 (-4.9)	0.895 (-2.2)	0.868 (-2.0)	0.778 (-2.9)	0.969 (-0.7)
P3: No state conditioning	0.932 (-1.0)	0.915 (-2.0)	0.720 (-3.9)	0.900 (-1.7)	0.872 (-1.6)	0.784 (-2.3)	0.970 (-0.6)
P4: No gate	0.925 (-1.7)	0.905 (-3.0)	0.700 (-5.9)	0.885 (-3.2)	0.860 (-2.8)	0.760 (-4.7)	0.965 (-1.1)
RoGRE (Full)	0.942	0.935	0.759	0.917	0.888	0.807	0.976

Design definitions. A0 contains the Initial Localization Proposal only. A1 adds the SAM2 dense-prompt decoder to A0, without Spatial Hypothesis Correction or Forensic Residual Refinement. A2 and A3 add Spatial Hypothesis Correction or Forensic Residual Refinement to A1, respectively. Parallel updates let SAM and LFFE update $\mathbf{z}^{(0)}$ independently; forensic→spatial reverses the full model's role order. State-agnostic gates remove $\mathbf{p}_t, \mathbf{u}_t$ from both gates, and ungated corrections use $\Delta_t = \tilde{\Delta}_t$.

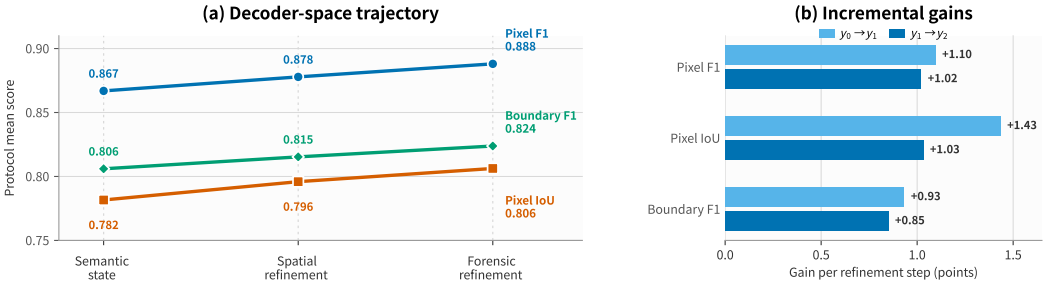


Figure 5: **Localization-State Evolution.** The same SAM2 decoder processes the Initial Localization Proposal (y_0), the state after Spatial Hypothesis Correction (y_1), and the state after Forensic Residual Refinement (y_2). Here, y_i is the output obtained by decoding the corresponding localization state $\mathbf{z}^{(i)}$. Panel (a) reports pixel F1, pixel IoU, and boundary F1 along the gated residual evolution trajectory; panel (b) reports the gain from each update in percentage points.

after Forensic Residual Refinement. For this diagnostic, the SAM2 decoder is fixed from the full RoGRE model, so its scores are not directly comparable to the separately trained A1 ablation. This decoder-controlled comparison shows that each successive state improves the final prediction, complementing the component and ordering ablations in Table 2.

5 Conclusion

RoGRE realizes controlled hypothesis revision by evolving a shared estimate across distinct cue roles. DINOv3 generates the Initial Localization Proposal, SAM2 image features perform Spatial Hypothesis Correction, and Bayar–SRM fingerprints conduct Forensic Residual Refinement. Bounded, state-conditioned gates regulate the latter two updates before the evolved state is decoded into the final mask. RoGRE achieves the highest mean pixel-level F1 among the reported methods under both evaluation protocols, while component, ordering, and gating ablations substantiate the three-stage design. Overall, these results support role-structured integration of heterogeneous cues without requiring each representation to produce an independent mask prediction.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing and debugging the experimental code used to train and evaluate RoGRE, including its Initial Localization Proposal, Spatial Hypothesis Correction, Forensic Residual Refinement, and ablation variants. We also used generative AI tools to organize and revise the prose describing the method and the reported experimental analyses. We did not use generative AI tools to generate synthetic datasets, collect data, or produce the reported experimental measurements. The authors executed the reported training and evaluation runs and manually checked all AI-assisted text against the final experimental results, tables, and citations. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops methods for image manipulation localization and evaluates them on established public benchmarks. It does not involve human-subject studies or the collection of new personal data. The method is intended to support forensic analysis of digital images; its outputs should be treated as auxiliary evidence rather than conclusive judgments about individuals or image provenance. We have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We support reproducibility through the methodological and experimental specifications provided in this manuscript. Sections 3 and 4 describe the model, objectives, training configuration, datasets, evaluation protocols, and metrics. Appendix A documents the architecture and optimization settings, while Appendices B–G provide the loss definitions, ablations, and additional analyses. **We will release our code after the paper is accepted.**

REFERENCES

Ningning Bai, Xiaofeng Wang, Ruidong Han, Jianpeng Hou, Yihang Wang, and Shanmin Pang. PIM-Net: Progressive inconsistency mining network for image manipulation localization. *Pattern Recognition*, 159:111136, 2025. doi: 10.1016/j.patcog.2024.111136.

Belhassen Bayar and Matthew C. Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pp. 5–10, 2016. doi: 10.1145/2909827.2930786.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

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

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Mask2Former: Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022.

Davide Cozzolino and Luisa Verdoliva. Noiseprint: A CNN-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2020. doi: 10.1109/TIFS.2019.2916364.

Jing Dong, Wei Wang, and Tieniu Tan. CASIA image tampering detection evaluation database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, pp. 422–426, 2013. doi: 10.1109/ChinaSIP.2013.6625374.

Hany Farid. Image forgery detection. *IEEE Signal Processing Magazine*, 26(2):16–25, 2009. doi: 10.1109/MSP.2008.931079.

Jonathan G. Fiscus, Haiying Guan, Yooyoung Lee, Amy Yates, Andrew Delgado, Daniel F. Zhou, and Timothée N. Kheyrkhah. Medifor nimble challenge evaluation. National Institute of Standards and Technology, 2017.

Jessica Fridrich and Jan Kodovský. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. doi: 10.1109/TIFS.2012.2190402.

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.

Kun Guo, Gang Cao, Zijie Lou, Xianglin Huang, and Jiaoyun Liu. A lightweight and effective image tampering localization network with vision mamba. *IEEE Signal Processing Letters*, 32:2179–2183, 2025.

Jing Hao, Zhixin Zhang, Shicai Yang, Di Xie, and Shiliang Pu. TransForensics: Image forgery localization with dense self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15055–15064, 2021. doi: 10.1109/ICCV48922.2021.01478.

Yu-Feng Hsu and Shih-Fu Chang. Detecting image splicing using geometry invariants and camera characteristics consistency. In *2006 IEEE International Conference on Multimedia and Expo*, pp. 549–552, 2006. doi: 10.1109/ICME.2006.262447.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Wen Huang, Jiarui Yang, Tao Dai, Jiawei Li, Shaoxiong Zhan, Bin Wang, and Shu-Tao Xia. RelayFormer: A unified local-global attention framework for scalable image and video manipulation localization. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28831–28841, 2025. doi: 10.1109/CVPR52734.2025.02685.

Kaixiang Ji, Feng Chen, Xin Guo, Yadong Xu, Jian Wang, and Jingdong Chen. Uncertainty-guided learning for improving image manipulation detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22456–22465, 2023.

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

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

Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning JPEG compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022. doi: 10.1007/s11263-022-01617-5.

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

Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022. doi: 10.1109/TCSVT.2022.3189545.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Jan Lukáš, Jessica Fridrich, and Miroslav Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, 2006. doi: 10.1109/TIFS.2006.873602.

Xiaochen Ma, Bo Du, Zhuohang Jiang, Xia Du, Ahmed Y Al Hammadi, and Jizhe Zhou. Imlvit: Benchmarking image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*, 2023.

Ghazal Mazaheri, Niluthpol Chowdhury Mithun, Jawadul H. Bappy, and Amit K. Roy-Chowdhury. A skip connection architecture for localization of image manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 119–129, 2019.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision*, pp. 565–571, 2016. doi: 10.1109/3DV.2016.79.

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.

Alin C. Popescu and Hany Farid. Exposing digital forgeries by detecting traces of resampling. *IEEE Transactions on Signal Processing*, 53(2):758–767, 2005. doi: 10.1109/TSP.2004.839932.

Chenfan Qu, Yiwu Zhong, Chongyu Liu, Guitao Xu, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards modern image manipulation localization: A large-scale dataset and novel methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10781–10790, 2024.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7262–7272, 2021.

Lei Su, Xiaochen Ma, Xuekang Zhu, Chaoqun Niu, Zeyu Lei, and Ji-Zhe Zhou. Can we get rid of handcrafted feature extractors? SparseViT: Nonsemantics-centered, parameter-efficient image manipulation localization through sparse-coding transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7024–7032, 2025. doi: 10.1609/aaai.v39i7.32754.

Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2364–2373, 2022.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. OpenSDI: Spotting diffusion-generated images in the open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4291–4301, 2025.

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*, pp. 161–165, 2016. doi: 10.1109/ICIP.2016.7532339.

Yue Wu, Wael Abd-Almageed, and Prem Natarajan. BusterNet: Detecting copy-move image forgery with source/target localization. In *Proceedings of the European Conference on Computer Vision*, pp. 168–184, 2018.

Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9543–9552, 2019.

Yunyang Xiong, Bala Varadarajan, Lemeng Wu, Xiaoyu Xiang, Fanyi Xiao, Chenchen Zhu, Xiaoliang Dai, Dilin Wang, Fei Sun, Forrest Iandola, Raghuraman Krishnamoorthi, and Vikas Chandra. EfficientSAM: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16111–16121, 2024.

Jieming Yu, Qiuxiao Feng, Zhuohan Wang, and Xiaochen Ma. Dinov3 beats specialized detectors: A simple foundation model baseline for image forensics, 2026a.

Yanpu Yu, Zhaoxin Shi, Hanqing Zhao, Tianyi Wei, Wenbo Zhou, and Nenghai Yu. EARG-Net: Edge-aware reconstruction-guided network for image manipulation detection and localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12178–12186, 2026b. doi: 10.1609/aaai.v40i14.38208.

Jiwei Zhang, Wenbo Feng, Siwei Wang, Feifei Kou, Haoyang Yu, and Shaozhang Niu. Collaborative transformers with multi-level forensic attention for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 12556–12563, 2026. doi: 10.1609/aaai.v40i15.38250.

Quan Zhang, Yuxin Qi, Xi Tang, Jinwei Fang, Xi Lin, Ke Zhang, and Chun Yuan. IMDPrompter: Adapting SAM to image manipulation detection by cross-view automated prompt learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jizhe Zhou, Xiaochen Ma, Xia Du, Ahmed Y. Alhammadi, and Wentao Feng. Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22346–22356, 2023.

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.

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*, pp. 11022–11030, 2025. doi: 10.1609/aaai.v39i10.33198.

Xuekang Zhu, Ji-Zhe Zhou, Kaiwen Feng, Chenfan Qu, Xiwen Wang, Yunfei Wang, Liting Zhou, and Jian Liu. Revisiting image manipulation localization under realistic manipulation scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, 2026.

Appendix for RoGRE

A Architecture and Training Details

Backbone Adaptation and Prompt Resolution. All images are resized to 512×512. We use DINOv3 ViT-L/16 (Siméoni et al., 2025) as the semantic encoder and SAM2 Hiera-B+ (Ravi et al., 2025) as the spatial encoder. Both image encoders receive rank-32 LoRA adapters (Hu et al., 2022) with scaling factor 64 in their query, key, and value projections. All pretrained encoder weights are frozen apart from the LoRA parameters; the SAM2 prompt encoder and mask decoder remain trainable. The DINOv3 locator consists of two convolution–group-normalization–activation blocks and a 1×1 logit projection. At each forward pass, the dense-prompt lattice is set to four times the spatial resolution of the SAM2 image embedding, matching the mask-input resolution expected by the SAM2 prompt encoder. This common lattice lets the Initial Localization Proposal, Spatial Hypothesis Correction, and Forensic Residual Refinement operate on the same localization state before final decoding.

Gated Residual Correction Heads. The SAM2 correction head projects the main SAM2 embedding and its two high-resolution features to 64, 32, and 32 channels, respectively. Their concatenation is fused into a 128-channel spatial context. LFFE yields a 64-channel forensic fingerprint, which is resized bilinearly and mapped to a separate 128-channel forensic context. Each branch applies the gated residual rule in Section 3.3 to its incoming state, with a signed candidate range of $[-2, 2]$. The forensic residual head initializes both its candidate projection and final gate projection to zero. Its initial candidate and applied update are therefore zero, so optimization starts from the semantic–spatial pathway.

Forensic Fingerprint Encoder. Before residual extraction, LFFE reverses ImageNet normalization and clips the input RGB values to $[0, 1]$. Its Bayar bank follows constrained prediction-error filtering (Bayar & Stamm, 2016) with twelve grouped 5×5 filters, each fixing the center coefficient to -1 and projecting the remaining 24 coefficients to sum to one. In parallel, a twelve-channel learnable 5×5 SRM bank is initialized from canonical residual kernels (Fridrich & Kodovský, 2012) combined with R-, G-, B-, and luminance-specific mixing vectors. We remove the spatial mean of every learned SRM kernel at each forward pass to retain a zero-DC response. The two residual banks are independently projected to 16 channels and fused into a 32-channel stem. A local path combines 3×3, 5×5, and dilated 3×3 depthwise receptive fields at half resolution with a quarter-resolution context path, producing the 64-channel fingerprint. Both residual-filter banks are evaluated in FP32 when the remaining network uses automatic mixed precision, and LFFE has no raw-RGB bypass.

Optimization and Augmentation. We train for 50 epochs using AdamW (Loshchilov & Hutter, 2019) with learning rate 10^{-4} and weight decay 0.05. The learning rate linearly warms up for the first five epochs and remains constant afterwards. The batch size is eight with no gradient accumulation. During training, cross-image splice augmentation is applied with probability 0.1. The reported configuration uses the same model, loss, and augmentation settings for all experiments.

A.1 Hyperparameter Configuration

Table 3 consolidates all model, loss, optimization, and inference hyperparameters specified for the reported configuration. Unless a setting is varied explicitly in an ablation, the same values are used throughout the experiments.

B Loss Details

All targets are resized by nearest-neighbor interpolation to the resolution of the supervised logits. For a pixel logit z_i, binary target $y_i \in \{0, 1\}$, and sigmoid $\sigma(\cdot)$, the weighted binary cross-entropy used in $\mathcal{L}_{\text{mask}}$ is computed elementwise as

$$\ell_{\text{wbce}}(z_i, y_i) = -\lambda_1 y_i \log \sigma(z_i) - \lambda_2 (1 - y_i) \log(1 - \sigma(z_i)), \tag{7}$$

where $\lambda_1 = 2$ and $\lambda_2 = 0.5$.

Table 3: **RoGRE Hyperparameter Configuration.** Model, gated residual evolution, objective, optimization, and inference settings used throughout the reported experiments unless explicitly varied in an ablation.

Model and prompt		Loss, optimization, and inference	
Input resolution	512×512	Mask-loss weights $(\mathcal{L}_{\text{wbce}}, \mathcal{L}_{\text{dice}})$	$(0.2, 0.8)$
Semantic / spatial backbone	DINOv3 ViT-L/16 / SAM2 Hiera-B+	Weighted-BCE weights (λ_1, λ_2)	$(2, 0.5)$
LoRA target projections	query, key, value	Dice smoothing ϵ	10^{-4}
LoRA rank / scaling	32 / 64	Boundary-gradient kernel size	7
DINO locator blocks	2	Forensic FN / FP probability thresholds	0.5 / 0.7
SAM2 projection / context channels	64, 32, 32 / 128	Auxiliary weights $(\lambda_D, \lambda_S, \lambda_F)$	$(0.1, 0.1, 0.1)$
Bayar / SRM filters	12 / 12 (5×5)	Optimizer	AdamW
Residual-bank projection channels	16 per bank	Learning rate / weight decay	10^{-4} / 0.05
LFFE fingerprint channels	64	Training epochs / warm-up epochs	50 / 5
Candidate range $[-\alpha_t, \alpha_t]$	$[-2, 2]$ for $t \in \{\text{S}, \text{F}\}$	Batch size / gradient accumulation	8 / none
Dense-prompt lattice	$4\times$ SAM2 embedding resolution	Cross-image splice probability	0.1
Forensic candidate / gate initialization	zero / zero	Evaluation threshold	0.5

For a mini-batch $\mathcal{B}$, let $\mathcal{B}_+ \subseteq \mathcal{B}$ denote the samples containing manipulated foreground, and let $\epsilon = 10^{-4}$. The squared Dice term in equation 5 is

$$\mathcal{L}_{\text{dice}}(\mathbf{z}, \mathbf{y}) = \frac{1}{|\mathcal{B}_+|} \sum_{n \in \mathcal{B}_+} \left[1 - \frac{2 \sum_i \sigma(z_{n,i}) y_{n,i} + \epsilon}{\sum_i \sigma(z_{n,i})^2 + \sum_i y_{n,i}^2 + \epsilon} \right]. \tag{8}$$

If $\mathcal{B}_+$ is empty, we set $\mathcal{L}_{\text{dice}}$ to zero. Consequently, authentic images are supervised only by weighted binary cross-entropy. Equation equation 5 applies this mask objective to both the final decoder output and the initial DINOv3 proposal.

Spatial Boundary Loss. The spatial state $\mathbf{z}^{(1)}$ is supervised on a boundary support obtained from a morphological gradient with kernel size seven:

$$\mathbf{b} = \text{clip}\big(\text{MaxPool}_7(\mathbf{y}) - \big[-\text{MaxPool}_7(-\mathbf{y})\big], 0, 1\big)\,, \tag{9}$$

where MaxPool_7 is max pooling with kernel size seven, unit stride, and padding three. Equivalently, $\mathbf{b}$ is the dilation of $\mathbf{y}$ minus its erosion. For each sample with nonempty boundary support, we normalize the weighted BCE by $\sum_i b_i$. We compute the squared Dice in the same way as equation 8 after replacing $\sigma(z_i)$ and y_i with $b_i\sigma(z_i)$ and $b_i y_i$. Writing the two terms as $\mathcal{L}_{\text{bdry-wbce}}$ and $\mathcal{L}_{\text{bdry-dice}}$, the boundary objective is

$$\mathcal{L}_{\text{bdry}} = \frac{1}{2}(\mathcal{L}_{\text{bdry-wbce}} + \mathcal{L}_{\text{bdry-dice}})\,, \tag{10}$$

which is evaluated on the resolution-matched spatial state $\mathbf{z}^{(1)}$.

Forensic Residual Loss. The forensic loss supervises corrections to the incoming spatial state rather than an independent mask. We derive two complementary residual-error supports from the detached, resolution-matched probability $\overline{\mathbf{p}}^{(1)} = \sigma(\text{sg}(\mathbf{z}^{(1)}))$:

$$\mathbf{w}_{\text{fn}} = \mathbf{y} \odot \text{clip}\left(\frac{0.5 - \overline{\mathbf{p}}^{(1)}}{0.5}, 0, 1 \right),$$
$$\mathbf{w}_{\text{fp}} = (1 - \mathbf{y}) \odot \text{clip}\left(\frac{\overline{\mathbf{p}}^{(1)} - 0.7}{0.3}, 0, 1 \right). \tag{11}$$

The foreground support emphasizes missed manipulated pixels with incoming probability below 0.5, whereas the background support emphasizes false-positive pixels with incoming probability above 0.7. Let $\mathbf{z}_{\text{F}} = \text{sg}(\mathbf{z}^{(1)}) + \Delta_{\text{F}}$ denote the forensic-corrected logits. For sample n, support $\mathbf{w}_n$, target $\mathbf{t}_n$, and candidate region $\mathbf{c}_n$, we define

$$\mathcal{R}_n(\mathbf{z}_{\text{F}}, \mathbf{t}; \mathbf{w}, \mathbf{c}) = \frac{\sum_i w_{n,i}}{\sum_i c_{n,i} + \epsilon} \frac{\sum_i w_{n,i} \ell_{\text{wbce}}(z_{\text{F},n,i}, t_{n,i})}{\sum_i w_{n,i} + \epsilon}. \tag{12}$$

Table 4: **Role-Aligned Objective Weight Ablation.** Pixel-level F1 and four-dataset mean metrics under the CAT-Net protocol for different weights on Initial Localization Proposal, Spatial Hypothesis Correction, and Forensic Residual Refinement.

Variant	Auxiliary-loss weights			Pixel-level F1				Four-dataset mean		
	λ_D	λ_S	λ_F	CASIAv1	Columbia	NIST16	Coverage	F1	IoU	Acc.
Default	0.10	0.10	0.10	0.942	0.935	0.759	0.917	**0.888**	**0.807**	**0.976**
A1: $\lambda_D = 0.00$	0.00	0.10	0.10	0.934	0.928	0.748	0.907	0.880	0.797	0.973
A2: $\lambda_D = 0.20$	0.20	0.10	0.10	0.938	0.931	0.750	0.910	0.882	0.800	0.974
A3: $\lambda_S = 0.00$	0.10	0.00	0.10	0.932	0.926	0.746	0.905	0.879	0.796	0.973
A4: $\lambda_S = 0.20$	0.10	0.20	0.10	0.936	0.929	0.748	0.908	0.881	0.799	0.974
A5: $\lambda_F = 0.00$	0.10	0.10	0.00	0.932	0.925	0.742	0.906	0.876	0.792	0.973
A6: $\lambda_F = 0.20$	0.10	0.10	0.20	0.934	0.927	0.744	0.907	0.878	0.795	0.973
A7: no role auxiliaries	0.00	0.00	0.00	0.925	0.915	0.730	0.895	0.870	0.783	0.970

where i indexes pixels and the first factor measures the remaining-error ratio within the candidate region. Averaging only over samples with nonempty support gives

$$\mathcal{L}_{\text{for}} = \underset{n:\|\mathbf{w}_{\text{fn},n}\|_1>0}{\text{mean}} \mathcal{R}_n(\mathbf{z}_{\text{F}}, \mathbf{y}; \mathbf{w}_{\text{fn}}, \mathbf{y}) + \underset{n:\|\mathbf{w}_{\text{fp},n}\|_1>0}{\text{mean}} \mathcal{R}_n(\mathbf{z}_{\text{F}}, \mathbf{y}; \mathbf{w}_{\text{fp}}, 1-\mathbf{y}). \tag{13}$$

With $\lambda_D = \lambda_S = \lambda_F = 0.1$, the complete objective takes the same form as equation 6:

$$\boxed{\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mask}}(\mathbf{z}^{\text{out}}, \mathbf{y}) + \lambda_D \mathcal{L}_{\text{mask}}(\mathbf{z}^{(0)}, \mathbf{y}) + \lambda_S \mathcal{L}_{\text{bdry}}(\mathbf{z}^{(1)}, \mathbf{y}) + \lambda_F \mathcal{L}_{\text{for}}.} \tag{14}$$

The final term supervises the signed update from Forensic Residual Refinement over the residual-error supports.

C Loss-Weight Ablation

Table 4 examines the three auxiliary terms in equation 6: λ_D supervises the Initial Localization Proposal, λ_S supervises Spatial Hypothesis Correction at manipulation boundaries, and λ_F supervises Forensic Residual Refinement on residual errors. The full configuration assigns 0.10 to each auxiliary term and attains mean F1/IoU/accuracy of $0.888/0.807/0.976$.

Among the reported alternatives, setting λ_D, λ_S, or λ_F to zero reduces the mean F1 by 0.008, 0.009, and 0.012, respectively. Increasing any one weight to 0.20 also lowers mean F1, to 0.882, 0.881, and 0.878. The forensic term has the largest individual effect in the table, particularly on NIST16, where removing it reduces F1 from 0.759 to 0.742. Removing all three role-aligned auxiliaries further lowers mean F1, IoU, and accuracy to 0.870, 0.783, and 0.970. These results support the balanced 0.10 weighting: each auxiliary loss contributes, while overemphasizing a single stage weakens the final localization objective.

D State Evolution and Final Mask Decoding

Diagnostic Setup. Our method first evolves a shared localization state through three cue roles and then decodes the final state into a mask. To disentangle state evolution from final mask decoding, we evaluate the Initial Localization Proposal $\mathbf{z}^{(0)}$, the state after Spatial Hypothesis Correction $\mathbf{z}^{(1)}$, the state after Forensic Residual Refinement $\mathbf{z}^{(2)}$, and the final decoder output $\mathbf{z}^{\text{out}}$ at a threshold of 0.5. We analyze all 3,437 CAT-Net test images, including 1,673 authentic images. The all-image F1 in this diagnostic includes both tampered and authentic images, whereas the main results average dataset-level F1 computed only on tampered images. The full test list also allows false-positive rate (FPR) on authentic images to be evaluated alongside manipulation localization. Table 5 quantifies the error conversion induced by each transition, while Figure 6 summarizes foreground support, FPR, and localization quality. The first three maps are diagnostic intermediate states rather than independently supervised final masks.

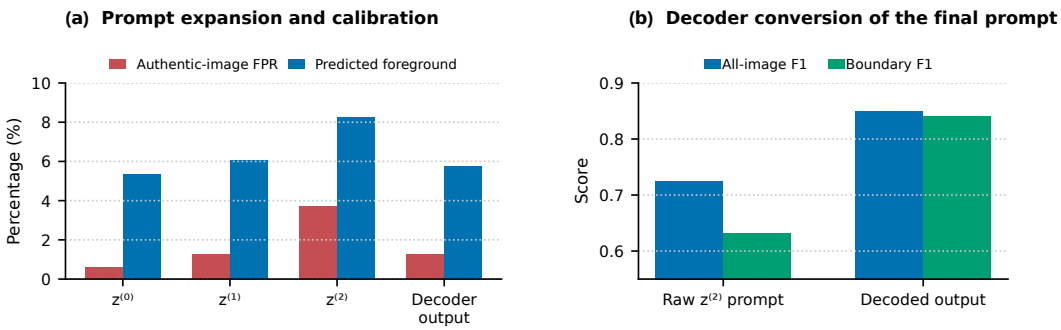


Figure 6: **Gated Residual Evolution and Final Mask Decoding.** (a) Spatial Hypothesis Correction and Forensic Residual Refinement progressively expand foreground support, accompanied by increased false-positive rates on authentic images. (b) Final mask decoding sharply suppresses false positives while improving all-image F1 (including tampered and authentic images) and boundary F1. Figure 5 separately reports the localization-state trajectory under a shared decoder.

Table 5: **Error Conversion Across Gated Residual Evolution and Final Mask Decoding.** Pooled CAT-Net test-list results at threshold 0.5. All entries are percentages: FN recovery is relative to incoming FNs, FP suppression is relative to incoming FPs, and new errors are relative to previously correct pixels. Spatial Hypothesis Correction and Forensic Residual Refinement recover missed foreground, while final mask decoding suppresses most false positives in the evolved state.

Space	Transition	FN recovery (%)	FP suppression (%)	New errors (%)
State update	$\mathbf{z}^{(0)} \to \mathbf{z}^{(1)}$ (SAM)	22.85	2.39	0.56
State update	$\mathbf{z}^{(1)} \to \mathbf{z}^{(2)}$ (LFFE)	30.39	1.98	2.09
Decoder	$\mathbf{z}^{(2)} \to \mathbf{z}^{\text{out}}$	0.25	69.04	0.25

Role-Specific Updates Recover Missed Regions. The two gated residual updates primarily recover manipulated regions missed by the incoming state. Spatial Hypothesis Correction, $\mathbf{z}^{(0)} \to \mathbf{z}^{(1)}$, recovers 22.85% of the incoming false negatives, while suppressing only 2.39% of the incoming false positives. Forensic Residual Refinement, $\mathbf{z}^{(1)} \to \mathbf{z}^{(2)}$, further recovers 30.39% of the false negatives entering that stage, with only 1.98% false-positive suppression. The latter update is more aggressive, yielding greater FN recovery but also a higher new-error rate (2.09% versus 0.56%). Across the two updates, predicted foreground expands from 5.32% to 8.25%, while authentic-image FPR rises from 0.60% to 3.69%. These results indicate that state evolution primarily broadens candidate foreground support by accumulating complementary evidence, rather than producing a calibrated final mask.

Final Decoding Filters the Evolved State. Final mask decoding exhibits the complementary error-conversion behavior. Relative to $\mathbf{z}^{(2)}$, the SAM2 decoder recovers only 0.25% of the remaining false negatives but suppresses 69.04% of the incoming false positives, with a new-error rate of only 0.25%. Consequently, predicted foreground decreases from 8.25% to 5.73%, and authentic-image FPR drops from 3.69% to 1.26%. At the same time, all-image F1 increases from 0.724 to 0.850, while boundary F1 improves from 0.631 to 0.841. Thus, the gain from final decoding arises primarily from filtering false-positive support rather than recovering additional manipulated pixels. Together, these diagnostics reveal a clear functional division: gated state evolution accumulates candidate evidence from complementary cue roles, whereas image-conditioned decoding consolidates the evolved state into a more precise and spatially coherent final mask.

E Threshold Sensitivity and Mask-Area Analysis

Figure 7 summarizes the threshold sweep and a macro per-image analysis stratified by ground-truth mask area: small ($\leq 1\%$), medium (1–10%), and large ($> 10\%$). The fixed threshold 0.5 is near-optimal on average, achieving a mean pixel F1 of 0.888 versus 0.899 with independently tuned dataset-specific thresholds. Columbia has the largest gap, 0.024, whereas Coverage is maximized at 0.5. Relative to the Initial Localization Proposal, the final output improves macro per-image F1

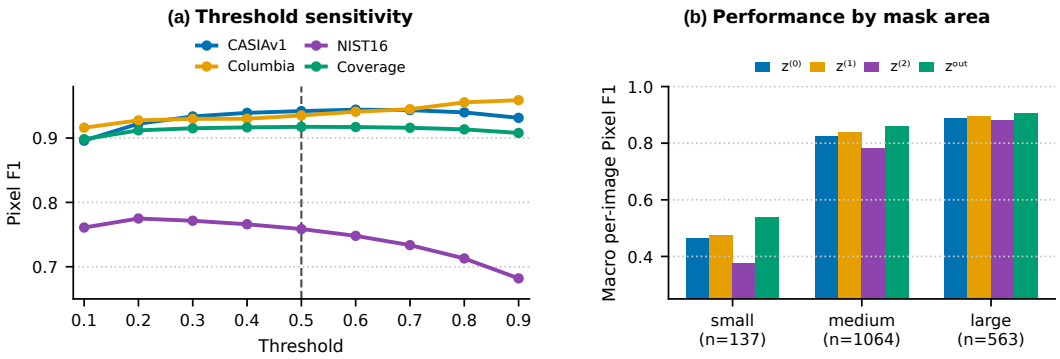


Figure 7: **Threshold Sensitivity and Mask-Area Analysis.** Panel (a) reports pixel-level F1 across decision thresholds; the dashed line marks the common threshold of 0.5. Panel (b) reports macro per-image pixel F1 for small, medium, and large ground-truth masks. $\mathbf{z}^{\text{out}}$ denotes the final mask-decoding output.

Table 6: **RoGRE Inference Cost.** Parameters, GMACs, latency, and peak allocated GPU memory at 512×512 resolution, batch size one, and FP32 precision.

Variant	Params (M)	Trainable (M)	GMACs	Latency (ms)	Peak memory (MB)
A1: +SAM2 Decoder	383.93	11.67	394.4	58.99	1710
A2: +Spatial correction	384.51	12.25	401.5	61.81	1714
A3: +LFFE	384.43	12.17	410.9	63.45	1768
Full	385.02	12.76	418.0	66.28	1778

by 0.073, 0.038, and 0.016 for small, medium, and large masks, respectively. The largest measured gain occurs for small manipulated regions; this comparison reflects the full gated residual evolution and decoding trajectory rather than any one stage in isolation.

F Compute and Inference Cost

We profile four RoGRE ablation variants on the same GPU using 512×512 inputs and FP32 inference. Table 6 shows that the full model adds 1.09M parameters and 23.5 GMACs over A1, which contains the DINOv3 semantic locator and SAM2 dense-prompt decoder but no correction stage. Mean latency increases by 7.29ms (12.4%), and peak allocated memory increases by 67MB (4.0%). MACs are estimated from torch-profiler-supported operators and should therefore be interpreted as a relative cost measure across variants rather than a hardware-independent analytical count.

G SID-Set Evaluation Across Image Origins

The main paper evaluates traditional local-manipulation benchmarks, whereas SID-Set (Huang et al., 2025) evaluates three AIGC-era image origins: AIGC-tampered images with localized edits, fully synthetic images, and real images. Table 7 reports both the original zero-shot setting and a SID-Set-trained setting. The AIGC-tampered subset uses local masks to assess manipulation localization, the full-synthetic subset uses all-foreground masks to assess image-wide synthetic-content recognition, and the real subset measures false-positive activations. In the SID-Set-trained setting, RoGRE is trained on the SID-Set training split and evaluated on the held-out validation split, while all other rows use no SID-Set training. At the image level, fake images are distinguished from real images using the 95th percentile of the predicted probability map.

Metrics and Their Roles. For the AIGC-tampered subset, we use pixel-level localization metrics. Prec. measures the fraction of predicted foreground pixels that correspond to annotated edited regions, whereas Rec. measures the fraction of annotated edited pixels recovered by the prediction. F1 summarizes the balance between precision and recall, while IoU measures the spatial overlap between the predicted and ground-truth masks.

Table 7: **SID-Set Evaluation Across Image Origins.** Tampered images use the official local masks. For full-synthetic images, every pixel is treated as foreground; for real images, FPR is the pixel false-positive rate (↓). RoGRE (zero-shot) and all comparison methods use no SID-Set training, whereas RoGRE (SID-Set-trained) is trained on the SID-Set training split and evaluated on the held-out validation split. Image-level $\mathrm{AUROC}_{\mathrm{p95}}$ and $\mathrm{AP}_{\mathrm{p95}}$ separate any fake image from real images using the 95th percentile of the predicted probability map. Best values are in **bold**.

Method	Tampered localization ↑				Full synthetic ↑		Real ↓	Any fake vs. real ↑	
	Prec.	Rec.	F1	IoU	FG Rec.	F1	FPR	$\mathrm{AUROC}_{\mathrm{p95}}$	$\mathrm{AP}_{\mathrm{p95}}$
CAT-Net (Kwon et al., 2022)	0.266	0.214	0.237	0.135	0.155	0.268	0.124	0.528	0.673
PSCC-Net (Liu et al., 2022)	0.199	0.357	0.256	0.146	0.251	0.401	0.229	0.552	0.690
MVSS-Net (Chen et al., 2021)	0.227	0.023	0.042	0.022	0.016	0.032	0.016	0.501	0.668
TruFor (Guillaro et al., 2023)	0.398	0.082	0.136	0.073	0.026	0.051	0.033	0.450	0.637
Mesorch (Zhu et al., 2025)	0.724	0.128	0.217	0.122	0.003	0.006	0.009	0.477	0.672
RITA (Zhu et al., 2026)	0.543	0.517	0.530	0.360	0.105	0.191	0.147	0.461	0.741
Relay-ViT (Huang et al., 2026)	0.253	0.019	0.036	0.018	0.006	0.011	0.009	0.456	0.645
DINOv3-IML (Yu et al., 2026a)	0.761	0.272	0.420	0.251	0.003	0.010	0.004	0.527	0.689
RoGRE (zero-shot)	0.828	0.392	0.532	0.363	0.007	0.013	0.010	0.602	0.777
RoGRE (SID-Set-trained)	**0.962**	**0.966**	**0.964**	**0.931**	**0.999**	**0.999**	**0.006**	**0.940**	**0.974**

For full-synthetic images, the target is $y_i = 1$ at every pixel. FG Rec. therefore measures the fraction of pixels recognized as synthetic at the fixed threshold of 0.5, testing whether the model covers an entire generated image rather than only a localized AIGC edit. Conversely, real images have $y_i = 0$ everywhere, and FPR quantifies spurious pixel-level forgery activations. Together, the three subsets distinguish localized AIGC manipulation, globally generated content, and authentic imagery, which are not separated by the traditional local-manipulation benchmarks used in the main paper.

For image-level evaluation, each image x receives the score $s(x) = Q_{0.95}(\{p_i\}_i)$, i.e., the 95th percentile of its predicted probability map. $\mathrm{AUROC}_{\mathrm{p95}}$ measures whether fake images receive higher percentile scores than real images across all decision thresholds, while $\mathrm{AP}_{\mathrm{p95}}$ summarizes the corresponding precision–recall curve. Compared with directly using the maximum response, the percentile-based score requires spatially supported high-confidence evidence and is therefore less sensitive to isolated activation peaks. The "any fake vs. real" protocol pools AIGC-tampered and full-synthetic images as the positive class, so these metrics jointly evaluate image-level discrimination across both localized and fully generated content.

Zero-Shot Performance. In the zero-shot setting, RoGRE transfers to the localized AIGC-tampered subset with an F1 score of 0.532 and an IoU of 0.363. Its precision is relatively high at 0.828, whereas its recall is 0.392, indicating that under the fixed operating threshold, the model tends to produce compact and high-confidence edited regions rather than aggressively covering uncertain areas. Despite not being trained on SID-Set, RoGRE also maintains a low pixel-level FPR of 0.010 on real images, suggesting that its localization responses remain well suppressed on authentic content.

The full-synthetic subset poses a different task from localized AIGC tampering: the entire image should be recognized as foreground rather than a spatially confined edit. Such image-wide targets are absent from RoGRE's original localized-manipulation training objective. Consequently, without explicit supervision for globally synthetic content, the model may produce weak or spatially incomplete responses instead of activating consistently over the whole image.

This distinction is also reflected in image-level discrimination. Using the same prediction maps, the zero-shot model obtains $\mathrm{AUROC}_{\mathrm{p95}} = 0.602$ and $\mathrm{AP}_{\mathrm{p95}} = 0.777$ for distinguishing any fake image from real images. Since this protocol treats both AIGC-tampered and full-synthetic images as positives, incomplete activation on unseen full-synthetic samples reduces the overall separability between fake and real images. The result nevertheless shows that the localized-manipulation model retains useful image-level discrimination under this broader AIGC setting.

Performance After SID-Set Training. After training on the SID-Set training split, RoGRE improves across all three image origins. For AIGC tampering, it reaches an F1 score of 0.964 and an IoU of 0.931, with well-balanced precision and recall of 0.962 and 0.966, respectively. These results show that SID-Set supervision improves mask coverage while preserving high localization precision.

More importantly, explicit exposure to full-synthetic samples enables the model to learn that globally generated images should induce foreground responses over the entire spatial extent. As a result, RoGRE achieves an F1 score of 0.999 on full-synthetic images, indicating that the distribution gap observed in the zero-shot setting can be effectively addressed once corresponding supervision is introduced. At the same time, the pixel-level FPR on real images further decreases from 0.010 to 0.006, showing that the increased sensitivity to synthetic content does not come at the cost of increased false activations on authentic images.

These improvements further translate into substantially stronger image-level discrimination. The SID-Set-trained model achieves $\mathrm{AUROC}_{\mathrm{p95}} = 0.940$ and $\mathrm{AP}_{\mathrm{p95}} = 0.974$ for the "any fake vs. real" setting. The large gain over the zero-shot model is consistent with the improved coverage of full-synthetic images: once both localized manipulations and globally generated content produce spatially supported high-confidence responses, the percentile-based image score can more reliably separate fake images from real ones.

Overall, SID-Set complements the main-paper evaluation by separating localized AIGC tampering, full-image synthesis, and real imagery. RoGRE transfers to localized AIGC tampering without SID-Set training, while full-synthetic recognition benefits from direct exposure to image-wide synthetic targets. With that supervision, the role-structured state evolution also supports these three image origins while retaining strong localized-manipulation performance.

H Scope and Limitations

RoGRE is designed for localized manipulation localization, where semantic context, spatial structure, and forensic residuals play complementary roles in revising a shared estimate. SID-Set extends this setting from traditional manipulation benchmarks to AIGC-tampered, full-synthetic, and real images. Zero-shot full-synthetic recognition and severe blur present different challenges: the former requires image-wide activation absent from the localized training objective, while the latter attenuates high-frequency traces available to the forensic stage. Table 7 shows that corresponding supervision closes the measured full-synthetic gap, and Figure 4 shows comparatively stable behavior under the tested noise and compression levels. These results define where the role-structured approach transfers and where its training target or input evidence must change.