CAN HETEROGENEOUS ARTIFACTS COEXIST? RGSA: ROLE-SPECIALIZED GLOBAL-SELECTIVE ADAPTATION FOR MITIGATING ARTIFACT FEATURE-SPACE COLLAPSE IN UNIFIED FAKE IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified Fake Image Detection (FID) aims to detect forgeries across four image forensic domains within a single model. However, joint multi-domain training substantially degrades the discriminative capability originally learned for individual domains, leading to a collapse of the artifact feature space. To uncover the underlying mechanism, we quantify heterogeneous conflicts during joint optimization and establish an evidence chain linking artifact heterogeneity to feature-space collapse. Our analysis reveals that cross-domain updates exhibit mixed and asymmetric interactions, indicating that different artifacts have distinct sharing scopes. We further find that organizing feature-space capacity according to these sharing scopes can effectively mitigate such collapse. Building on these findings, we propose Role-Specialized Global-Selective Adaptation (RGSA), which organizes the adaptation capacity into two pathways with distinct roles to accommodate different sharing scopes. The global pathway captures representations shared across all domains, whereas the selective pathway models partially shared and domain-specific artifacts. Under the four-domain unified protocol of OpenMMSec, RGSA effectively mitigates cross-domain conflicts and artifact feature-space collapse, and outperforms 23 state-of-the-art methods, validating both the proposed mechanism and the role-specialized design.

1 INTRODUCTION

Unified Fake Image Detection (FID) (Du et al., 2025) aims to perform unified real–fake discrimination across forensic domains, including **Deepfake** (Nguyen et al., 2019; Li et al., 2020b) (facial forgeries), **AIGC** (Ojha et al., 2023; Xi et al., 2023) (fully AI-generated images), **IMDL** (Zhou et al., 2023; Guillaro et al., 2023) (region-level manipulations in natural images), and **Doc** (Qu et al., 2023; Chen et al., 2024b) (document forgery), under realistic forensic scenarios where the underlying forgery method is unknown.

Model ensembles commonly address FID by routing inputs to domain-specific detectors (Huang et al., 2026a), but remain limited by routing errors, individual-detector bottlenecks, and costly domain-wise updates. A unified single model is therefore more desirable. SICA (Du et al., 2026) represents an initial step in this direction, while revealing that joint training degrades discriminability across all four forensic domains—a phenomenon termed artifact feature-space collapse. This degradation arises because domains differ substantially in generation mechanisms, frequency statistics, local structures, and semantic dependencies, resulting in heterogeneous dominant artifacts. The central challenge of unified FID is thus to **coordinate these artifacts without sacrificing the domain-discriminative representations required by each domain.**

Addressing this challenge requires further understanding how heterogeneous artifacts evolve into feature-space collapse during joint optimization. Heterogeneity itself does not necessarily induce harmful conflicts; only when updates driven by different artifacts continuously interfere with one another during joint optimization can they progressively undermine effective representations and eventually trigger feature-space collapse (Yu et al., 2020). We therefore further investigate: **How**

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.

severe are the conflicts among heterogeneous artifacts? How do these conflicts cause actual cross-domain harm and ultimately evolve into feature-space collapse?

We establish a three-level framework for tracing feature-space collapse during joint optimization. Gradient cosine similarity diagnoses potential directional conflicts, relative cross-domain harm quantifies their immediate impact against the target domain's learning scale, and single-versus-pairwise training comparisons assess whether this harm accumulates into long-term degradation. Together, they provide a quantitative chain from potential conflict to realized harm and negative transfer, making the formation of feature-space collapse observable.

Our diagnosis reveals mixed and asymmetric cross-domain interactions, ranging from beneficial to harmful transfer. This structure suggests that heterogeneous artifacts have different sharing scopes rather than a simple shared–private division: globally shared, partially shared across similar samples from certain domains, or domain-specific. Artifact feature-space collapse arises when incompatible artifact updates continually compete within a shared representation space.

Intuitively, expanding representational capacity may reduce competition among heterogeneous artifacts. However, our experiments show that capacity scaling alone does not consistently alleviate cross-domain harm or feature-space collapse. Under comparable capacity, reorganizing adaptation pathways yields better cross-domain interactions and representations, indicating that the key bottleneck lies in how capacity is aligned with artifact sharing scopes. Fully shared adaptation forces incompatible updates to compete, whereas strict domain-wise separation prevents reusable artifacts from being shared. Unified FID therefore requires structured capacity that supports stable global sharing, input-dependent partial sharing, and necessary specialization.

Based on the above analysis, we hypothesize that adaptation pathways should be specialized according to the cross-domain sharing scopes of artifacts. Following this principle, we propose **Role-Specialized Global-Selective Adaptation (RGSA)**, consisting of global and selective pathways. The global pathway combines OFT (Qiu et al., 2023) with relational anchoring to constrain cross-domain updates and stably capture globally shared artifacts. The selective pathway routes each input to LoRA (Hu et al., 2021a) experts, allowing similar artifacts to share experts while isolating domain-specific ones. Together, RGSA accommodates global, partial, and domain-specific sharing within a unified model.

We systematically validate the proposed mechanism and method on OpenMMSec (Du et al., 2026). The protocol uses 26 fine-grained forgery types for training and validation, and evaluates generalization on 72 unseen types spanning four forensic domains. Experiments show that RGSA reduces cross-domain harm, mitigates feature-space collapse under joint training, and substantially improves discriminative performance. RGSA demonstrates substantial performance advantages over a broad range of general-purpose backbones, domain-specific detectors, and unified detection methods. Further analyses across different backbones, pathway combinations, expert activation patterns, and isolated branches consistently support the role-specialization hypothesis.

In short, we systematically investigate how heterogeneous artifacts lead to feature-space collapse in unified FID. Our contributions are fourfold:

1. **Quantify the Formation of Feature-Space Collapse.** We trace its formation through three levels: directional conflict, relative cross-domain harm, and long-term negative transfer. We further reveal that feature-space collapse does not simply arise from insufficient capacity, but primarily from a mismatch between capacity organization and the distinct sharing scopes of heterogeneous artifacts.
2. **Reveal the Cross-Domain Sharing Structure of Heterogeneous Artifacts.** We uncover a mixed and asymmetric interaction structure among cross-domain updates, and further show that heterogeneous artifacts exhibit distinct cross-domain sharing scopes.
3. **Introduce the Role-Specialization Hypothesis and RGSA.** We propose assigning complementary update pathways according to the cross-domain sharing scopes of artifacts, and instantiate this principle with Role-Specialized Global-Selective Adaptation (RGSA), which organizes feature-space capacity through role-specific pathways.
4. **Deliver Rigorous Validation.** Under the four-domain unified protocol of OpenMMSec, RGSA effectively mitigates artifact feature-space collapse during joint training and achieves leading performance across 72 unseen forgery types. Extensive analyses further validate the effectiveness of the role-specialized design and support the proposed hypothesis.

2 Related Work

Existing Work on Subdomains. Image forensics has long been divided into four tasks: (1) Deepfake Detection (Sun et al., 2026; Nguyen et al., 2019), (2) AI-generated Image Detection (Dhakal et al., 2026; Wang et al., 2026), (3) Image Manipulation Detection and Localization (Chen et al., 2021; Guillaro et al., 2023), and (4) Document Manipulation Localization (Qu et al., 2023; Chen et al., 2024b). Although research within individual subdomains has progressed rapidly, each domain exhibits generalization limitations from a unified perspective, preventing direct application to FID. Details of these four domains can be found in Appendix B.

Existing Work on FID. ForensicHub (Du et al., 2025) established a benchmark spanning four forensic domains. UniShield (Huang et al., 2026a) combines domain-specific detectors through adaptive routing, while SICA (Du et al., 2026) addresses heterogeneous artifacts with a monolithic detector. DeFakerOne (Team, 2026) builds a multimodal-large-model-based framework for unified forgery detection and localization, and reports both cross-domain transfer and interference. However, cross-domain interactions during joint FID training remain insufficiently characterized. Our work systematically investigates these interactions.

Artifact Heterogeneity and Cross-Domain Update Interactions. Recent work separately highlights artifact diversity and optimization conflicts in forensic learning. CS-DFD (Liu et al., 2026) addresses conflicting updates in Deepfake detection, while Forensic-MoE (Fang et al., 2025) uses specialized experts to capture diverse synthetic-image traces. Beyond forensics, ETAP (Ayman et al., 2026) predicts joint-training gains using estimates of how task-specific parameter updates affect other tasks' losses. Yet how such interactions arise across forensic domains and affect unified FID training remains unclear.

3 Artifact Feature-Space Collapse: Quantification and Causes

3.1 Optimization Perspective on Feature-Space Collapse

In unified FID, joint training may erode domain-specific representations, yet the optimization process underlying this collapse remains unclear. We therefore trace how cross-domain updates interact using gradient cosine similarity, relative cross-domain harm, and long-term transfer.

3.2 Quantifying Cross-Domain Interactions

Let $\mathcal{D} = \{D_1, D_2, D_3, D_4\}$ denote the four forensic domains and θ_t the trainable parameters at step t.

3.2.1 Gradient Compatibility

Following the established gradient-conflict criterion (Yu et al., 2020), we use cosine similarity to characterize the local compatibility of domain-conditioned updates. At the same parameter state θ_t, we compute the gradients on batches from domains A and B, respectively:

$$g_{A,t} = \nabla_\theta L_A(\theta_t), \qquad g_{B,t} = \nabla_\theta L_B(\theta_t). \tag{1}$$

Their gradient cosine similarity is computed as

$$C_{A,B,t} = \frac{g_{A,t}^\top g_{B,t}}{\|g_{A,t}\|_2 \|g_{B,t}\|_2 + \epsilon}. \tag{2}$$

Negative $C_{A,B,t}$ indicates locally conflicting update directions. $C_{A,B,t}$ is symmetric and magnitude-free; it diagnoses potential conflict rather than realized directional harm.

3.2.2 Relative Cross-Domain Harm

Following prior work on directional task interactions (Fifty et al., 2021), we apply a fixed-norm virtual update from source domain A and measure its effect on target domain B's loss without modifying

the model parameters.

$$\theta'_{A,t} = \theta_t - \rho \frac{g_{A,t}}{\|g_{A,t}\|_2 + \epsilon}, \tag{3}$$

$$\delta_{A\to B,t} = L_B(\theta'_{A,t}) - L_B(\theta_t). \tag{4}$$

Positive $\delta_{A\to B,t}$ indicates immediate harm. To remove loss-scale and within-domain effects, we normalize it against an equal-norm update from a non-overlapping target-domain batch B':

$$R_{A\to B,t} = \frac{\delta_{A\to B,t} - \delta_{B'\to B,t}}{|\delta_{B'\to B,t}| + \epsilon}. \tag{5}$$

$R > 0$ therefore indicates excess cross-domain harm relative to the in-domain reference.

3.2.3 Long-Term Transfer Effect

Prior work attributes joint-training degradation to compromised shared representations or feature-space collapse (Du et al., 2026). To examine whether the proposed $R_{A\to B}$ captures such interactions, we compare a pairwise model trained on $A + B$ with a single-domain model trained on B:

$$\Delta M_{A\to B} = M_B^{A+B} - M_B^B, \tag{6}$$

Here M_B^B and M_B^{A+B} are the target-domain accuracy (ACC) values after single-domain and pairwise joint training, respectively; $\Delta M_{A\to B} < 0$ indicates long-term negative transfer. Its agreement with $R_{A\to B}$ tests whether immediate harm predicts this persistent performance loss.

3.3 Interaction Patterns and Sharing Boundaries

We compare larger LoRA ranks with routed configurations to distinguish capacity scaling from capacity organization, following the frozen ViT (Dosovitskiy, 2020) LoRA adaptation paradigm of SICA (Du et al., 2026), with DINOv3 (Siméoni et al., 2025). Training-process statistics and average ACC are reported in Tab. 1 and Tab. 2, with additional LoRA-rank results provided in Appendix D.1.

Table 1: Gradient cosine similarity C across domain pairs for each model.

Model	DF–AIGC	DF–IMDL	DF–Doc	AIGC–IMDL	AIGC–Doc	IMDL–Doc	**Mean**
LoRA ($r = 16$)	0.016	0.021	0.007	-0.040	-0.007	0.010	0.001
LoRA ($r = 64$)	0.015	0.017	0.005	-0.036	-0.005	0.009	0.001
LoRA+8 routers ($r = 16$)	0.013	0.022	0.006	-0.032	-0.009	0.014	0.002
Frozen+8 routers ($r = 16$)	0.020	0.024	0.003	-0.033	-0.001	0.004	0.003

Table 2: Relative cross-domain harm R for each model. In each domain-pair cell, the upper value follows the domain order in the heading and the lower value is the reverse direction.

Model	DF–AIGC	DF–IMDL	DF–Doc	AIGC–IMDL	AIGC–Doc	IMDL–Doc	**Mean**	ACC
LoRA ($r = 16$)	0.379 -0.002	-0.081 -0.035	0.136 0.048	0.352 0.797	0.188 0.556	0.141 0.014	0.208	0.854
LoRA ($r = 64$)	0.393 0.009	0.028 0.019	0.158 0.082	0.297 0.681	0.173 0.525	0.101 0.041	0.209	0.851
LoRA+8 routers ($r = 16$)	0.277 -0.038	-0.027 -0.023	0.165 0.053	0.312 0.700	0.147 0.442	0.177 0.072	0.188	0.859
Frozen+8 routers ($r = 16$)	0.221 -0.024	-0.070 0.008	0.145 0.050	0.336 0.735	0.008 0.456	0.285 0.054	0.184	0.859

Cross-domain interactions are neither uniformly beneficial nor antagonistic. C diagnoses geometric conflict, whereas R quantifies its realized directional impact. Increasing the LoRA rank leaves the mean C and R largely unchanged, while routing reduces mean harm from 0.208 to 0.188 or 0.184, indicating that organizing adaptation capacity is more effective than merely enlarging it. The resulting mixed and asymmetric structure suggests that artifact information should exhibit different scopes of domain sharing, which can be broadly categorized as globally shared, partially shared, and domain-specific.

3.4 Instantaneous Harm and Collapse Risk

Fig. 1 compares the relative cross-domain harm $R_{A\to B}$ with the long-term transfer effect $\Delta M_{A\to B}$ across all directed domain pairs under LoRA-Rank16. The two metrics agree in 10 of the 12 directions, indicating that R effectively identifies directions prone to long-term negative transfer. Notably, updates from IMDL to AIGC yield both the largest relative harm and the largest long-term degradation, while opposite directions within the same pair can differ substantially, confirming asymmetric interactions.

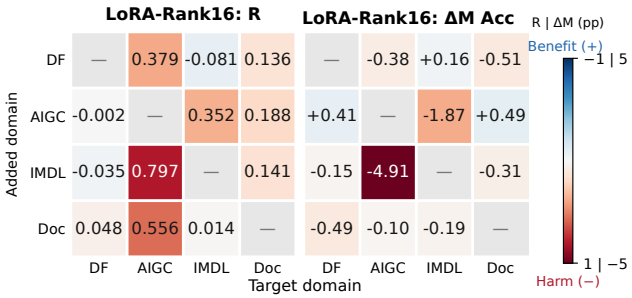

Figure 1: Relative cross-domain harm and long-term negative transfer across directed domain pairs.

Although R is not strictly monotonic with the final performance drop due to different baseline margins, their overall agreement indicates that local interference can accumulate into long-term degradation. Additional silhouette score (SIL) analysis is provided in Appendix D.3.

The overall consistency between local harm and long-term transfer effects suggests that heterogeneous interference can accumulate during joint optimization, progressively weakening the discriminative capability of the target domain. Therefore, R can serve as an indicator of the risk of local harm accumulating into long-term negative transfer and ultimately feature-space collapse.

3.5 Mechanistic Account of Feature-Space Collapse

Together, C, R, and ΔM form a three-level evidence chain: C diagnoses potential gradient conflict, R quantifies its immediate directional harm, and ΔM tests whether that harm persists as long-term degradation under joint training. Their overall consistency establishes a quantitative evidence chain from heterogeneous artifact interactions to feature-space collapse.

Artifact feature-space collapse is not directly caused by artifact heterogeneity or insufficient capacity. Rather, it arises from persistent competition among updates with different cross-domain sharing scopes when they are forced into an adaptation space without proper role allocation. Harmful interactions progressively weaken effective discriminative representations and eventually accumulate into feature-space collapse.

4 Role-Specialized Global-Selective Adaptation

4.1 Framework Overview

Motivated by the distinct artifact sharing scopes identified in Section 3, we propose Role-Specialized Global-Selective Adaptation (RGSA). As illustrated in Fig. 2, RGSA combines a global Orthogonal Finetuning (OFT) (Qiu et al., 2023) pathway with relational anchoring for globally shared artifacts and a routed LoRA pathway for partially shared and domain-specific artifacts.

4.2 Sharing Scopes and Role Specialization

For each sample, artifacts may be globally shared, partially shared, or domain-specific. If all components are written into a single update ΔW, partially shared and domain-specific artifacts may compete within the same parameter space. RGSA instead provides two complementary forms of adaptation capacity, global and selective:

$$h'(x) = W_0 h(x) + \Delta W_{\text{global}} h(x) + \sum_{e=1}^{E} p_e(x) \Delta W_e^{\text{res}} h(x). \tag{7}$$

Here, E is the number of experts, ΔW_{global} denotes the global adaptation update, ΔW_e^{res} the update of the e-th residual expert, and $p_e(x)$ the effective fusion weight after Top-k routing (Riquelme et al., 2021).

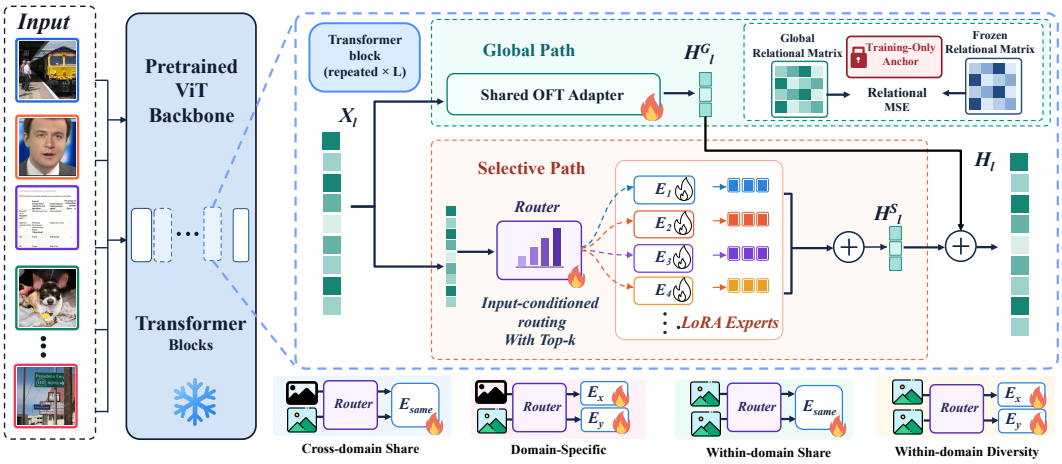

Figure 2: **Overview of RGSA.** Each adapted layer combines a globally shared OFT update with input-dependent residuals from Top-k LoRA experts. During training, a frozen reference branch provides relational anchors for the global pathway and is removed at inference.

Capacity separation alone is insufficient to induce functional specialization between pathways. RGSA therefore shapes complementary roles through differentiated parameterizations and relational anchoring.

4.3 Global Pathway: OFT and Relational Anchoring

Since the global pathway receives gradients from all domains, we constrain its feasible update space using OFT:

$$W_{\mathrm{O}} = W_0 Q, \qquad Q^\top Q = I. \tag{8}$$

Thus, $\Delta W_{\mathrm{global}} = W_0(Q - I)$ while $W_O W_O^\top = W_0 W_0^\top$, preserving the weight Gram structure. Since orthogonality alone does not constrain patch-relation drift, we further anchor the global representation to the frozen pretrained model.

We use the frozen pretrained model as a relational anchor. For the patch features $P^{(l)}$ from each selected layer l, we first center and normalize them to obtain $Z^{(l)}$, then define the relational matrix $\mathcal{C}^{(l)} = Z^{(l)} Z^{(l)\top}$.

Let $\mathcal{A}$ be the set of layers used to compute the anchor loss. The relational anchoring loss between the global pathway and the frozen teacher is defined as

$$\mathcal{L}_{\mathrm{anchor}} = \frac{1}{|\mathcal{A}|} \sum_{l \in \mathcal{A}} \mathrm{MSE}\left(\mathcal{C}^{(l)}_{\mathrm{global}}, \mathcal{C}^{(l)}_{\mathrm{teacher}}\right). \tag{9}$$

The frozen teacher receives no gradient. Relational anchoring permits feature reorganization while limiting patch-relation drift in the global-only branch; it does not constrain the selective experts and complements OFT's parameter constraint.

4.4 Selective Pathway: Routed LoRA Experts

The selective pathway activates independent LoRA experts through input-conditioned routing, providing capacity for partially shared and domain-specific artifacts that the global pathway may not accommodate. Cross-domain samples with similar artifacts can reuse experts, while differentiated activation can reduce competition among incompatible artifacts.

The e-th selective residual expert uses LoRA parameterization:

$$\Delta W_e^{\mathrm{res}} = \frac{s}{r} B_e A_e. \tag{10}$$

At adaptation layer l, let $u^{(l)}(x)$ denote that layer's input features. The router computes the pre-truncation probabilities $\pi^{(l)}(x) = \mathrm{softmax}(\mathrm{MLP}_l(u^{(l)}(x))/\tau)$, where τ is the routing softmax temperature.

Each adaptation layer retains the Top-k experts with the highest probability and renormalizes their weights to obtain $p^{(l)}(x)$; all other expert weights are zero. To balance expert usage and prevent persistent concentration on a few experts, we define

$$\mathcal{L}_{\mathrm{route}} = \lambda_{\mathrm{bal}} \frac{E}{|\mathcal{I}|} \sum_{l \in \mathcal{I}} \sum_{e=1}^{E} f_e^{(l)} \bar{\pi}_e^{(l)}. \tag{11}$$

Here, $\mathcal{I}$ is the set of adaptation layers; $f_e^{(l)}$ is expert e's share of all Top-k selections in the batch, measuring actual load without backpropagating through the discrete selections. The batch mean $\bar{\pi}_e^{(l)}$ uses pre-truncation probabilities, so even unselected experts receive a routing gradient. Assigning high probability to an already frequent expert increases the loss and encourages the router to redistribute usage.

4.5 Training Objective

The training objective is

$$\mathcal{L}_{\mathrm{train}} = \mathcal{L}_{\mathrm{BCE}}(\theta_g, \theta_s, \theta_r, \theta_h) + \mathcal{L}_{\mathrm{route}}(\theta_r) + \lambda_a \mathcal{L}_{\mathrm{anchor}}(\theta_g). \tag{12}$$

The trainable parameters $\theta_g, \theta_s, \theta_r, \theta_h$ belong to the global adapters, selective experts, router, and classification head, respectively. The pretrained backbone and teacher remain frozen. At inference, the frozen teacher and global-only auxiliary forward pass are removed.

5 Experiments

5.1 Experimental Setup

Data protocol. Following the standard evaluation protocol of the OpenMMSec benchmark (Du et al., 2026), 26 of its 98 fine-grained faking types are used for training and validation (80K training and 10K validation samples), while the remaining 72 types are reserved for testing (240K samples). We also report RGSA's performance on the Deepfake and AIGC subdomain benchmarks in Appendix G.3.

Implementation details. We adopt DINOv3 ViT-L/16 as the backbone. The selective pathway contains eight rank-16 experts, with Top-2 input-conditioned routing in each block and a router hidden dimension of 64.

Detectors. We compare with 23 methods, including general-purpose visual backbones, subdomain detectors, and unified FID detectors. The backbones are ResNet (He et al., 2016), EfficientNet (Tan & Le, 2019), SegFormer (Xie et al., 2021), Swin Transformer (Liu et al., 2021b), and ConvNeXt (Liu et al., 2022). Details of the subdomain and unified FID detectors are provided in Appendix C. We use ForensicHub (Du et al., 2025) as the codebase for implementation.

Cross-Domain Interaction Protocol. Independent probes compare domain updates at the same parameter state within each model without altering training. We fix domain sampling ratios, seeds, and data order across runs, then compute the metrics on each model's trainable parameters and average them over training epochs.

5.2 Generalization Comparison

We report results for each domain at the level of primary faking types. Tab. 3 reports accuracy (ACC), while the Appendix Tables 13, 14, and 15 report AUC, AP, and F1, respectively. We first compute the mean within each domain and then take the macro-average across domains as the overall score. RGSA achieves the highest overall score across all four metrics. In terms of ACC, it also obtains the best or joint-best domain average in all four domains.

Table 3: Accuracy (%) on OpenMMSec. The overall average is the macro-average of the four domain averages. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	Deepfake					AIGC							IMDL			Doc			Overall
	EFS	FE	FR	FS	Avg	GAN	Lat-Diff	Pix-Diff	AR	Comm	Other	Avg	Gen	Non-Gen	Avg	AIGC	Non-AIGC	Avg	Avg
ResNet	71.3	65.4	70.0	64.7	67.8	77.4	65.2	75.7	78.5	72.7	75.5	74.2	80.3	74.2	77.3	61.9	81.5	71.7	72.7
EfficientNet	45.0	29.0	50.8	48.5	43.3	65.2	58.9	63.1	65.1	62.1	63.7	63.0	51.7	51.1	51.4	47.0	76.8	61.9	54.9
CapsuleNet	63.4	58.7	66.5	59.6	62.0	79.2	63.1	77.5	81.3	73.2	77.8	75.4	79.2	75.3	77.2	59.8	84.8	72.3	71.7
SegFormer	83.8	80.5	83.5	75.1	80.7	87.4	80.9	87.1	89.7	87.8	82.3	85.9	83.2	80.3	81.7	66.8	80.1	73.4	80.4
Swin	83.6	80.5	80.6	71.5	79.0	86.9	83.8	85.8	87.6	86.8	81.3	85.4	85.0	80.8	82.9	66.9	77.1	72.0	79.8
ConvNeXt	80.5	84.0	83.9	72.4	80.2	89.0	79.8	87.9	92.0	89.6	83.4	87.0	83.6	80.2	81.9	65.0	75.3	70.1	79.8
Recce	59.0	50.8	50.7	53.7	53.5	90.4	66.1	90.8	97.1	85.3	87.8	86.3	63.1	77.1	70.1	78.6	52.8	65.7	68.9
Sia	64.0	63.5	73.1	63.7	66.1	81.6	68.7	80.2	84.1	79.2	80.4	79.0	71.9	63.8	67.9	64.4	83.4	73.9	71.7
IML-ViT	66.6	64.9	74.6	66.0	68.0	83.3	70.6	81.8	86.1	81.5	80.9	80.7	81.5	78.9	80.2	68.9	78.5	73.7	75.7
Trufor	73.7	72.7	77.1	65.4	72.2	83.7	77.9	83.0	86.0	84.0	80.3	82.5	82.7	78.3	80.5	58.1	86.6	72.3	76.9
UnivFD	57.0	44.8	57.6	57.4	54.2	84.1	68.3	81.7	84.9	76.9	82.9	79.8	65.7	76.2	70.9	39.0	86.3	62.7	66.9
FFDN	73.7	75.5	81.2	69.5	75.0	93.6	89.2	93.9	96.1	94.5	86.8	92.4	84.7	80.6	82.6	64.8	75.3	70.0	80.0
Effort	87.1	87.2	80.7	85.0	85.0	86.6	71.3	84.6	86.3	78.8	83.9	81.9	81.7	85.7	83.7	61.1	78.1	69.6	80.1
Mesorch	76.7	72.6	79.9	73.4	75.7	83.0	73.8	82.1	85.5	83.3	81.0	81.4	81.2	78.7	79.9	63.6	81.8	72.7	77.4
CO-SPY	75.1	71.3	77.0	65.5	72.2	84.7	80.8	84.2	86.2	85.7	78.5	83.3	77.8	74.2	76.0	61.9	78.0	70.0	75.4
AIDE	77.2	69.9	73.1	67.4	71.9	84.7	78.1	83.3	84.1	81.2	82.9	82.4	81.5	83.7	82.6	54.3	81.2	67.7	76.2
LEGION	83.7	82.4	79.5	78.5	81.0	91.0	78.0	89.2	91.1	88.2	83.2	86.8	83.0	86.1	84.6	68.6	79.3	**74.0**	81.6
PPMClip	86.5	82.8	81.3	76.5	81.8	93.6	89.5	92.1	93.6	90.4	82.4	90.3	84.0	83.6	83.8	63.8	80.5	72.2	82.0
SIMLBR	86.7	87.3	65.2	60.2	74.8	92.7	81.8	93.3	96.6	90.4	93.1	91.3	64.3	77.3	70.8	77.4	49.0	63.2	75.0
DFD-HR	91.4	84.5	86.6	83.7	86.5	97.2	95.6	96.0	97.2	95.1	85.7	94.5	85.1	83.8	84.5	67.6	79.7	73.6	84.8
RelayFormer	71.2	71.3	77.2	62.0	70.4	95.1	91.2	92.9	95.3	93.3	84.7	92.1	85.0	78.4	81.7	63.6	81.7	72.7	79.2
IAPL	88.3	81.9	78.3	80.3	82.2	89.1	88.4	87.4	88.7	85.6	78.1	86.2	84.9	84.4	84.6	67.0	75.4	71.2	81.1
SICA	91.3	89.4	86.5	86.6	88.4	96.3	94.5	94.7	96.7	94.8	86.9	94.0	85.4	85.1	85.3	69.2	78.4	73.8	85.4
RGSA (Ours)	93.9	93.3	88.3	87.9	**90.9**	96.3	94.8	95.5	96.2	94.8	90.5	**94.7**	86.4	85.5	**85.9**	69.3	78.8	**74.0**	**86.4**

5.3 Quantifying Heterogeneous Harm

Tab. 4 compares RGSA, LoRA, OFT, and OFT with routed LoRA using cross-domain interaction measures and average ACC. RGSA yields the lowest relative harm and the highest average ACC among these methods. Relative harm decreases from 0.208 for LoRA to 0.179 after introducing routed capacity, and further to 0.130 with RGSA. Additional pairwise interaction results are reported in Appendix Fig. 3 and Fig. 4.

Table 4: Quantifying Heterogeneous Harm and avg ACC

Model	COS	R	ACC	Model	COS	R	ACC
LoRA ($r = 16$)	0.001	0.208	0.854	OFT	0.002	0.191	0.855
OFT + routed LoRA	0.009	0.179	0.860	**RGSA**	0.006	**0.130**	**0.864**

5.4 Validating Feature-Space Collapse Mitigation

Tab. 5 validates whether RGSA mitigates feature-space collapse by comparing single-domain and unified training. The single-domain row combines four separately single-domain trained models, using only the results from their respective training domains, while the unified row reports the results of a single jointly trained model.

Unified RGSA raises mean ACC from 0.8615 to 0.8637, improving Deepfake and Doc while remaining competitive on AIGC and IMDL. These results demonstrate that RGSA mitigates feature-space collapse while preserving domain-discriminative performance under unified training. The corresponding feature-space visualization and extra SIL analysis are provided in Appendix E.

Table 5: ACC of unified RGSA versus its single-domain counterpart on each target domain. This single-domain approach combines separately trained models, utilizing only the results from the training domain; Mean is the macro-average across four domains.

Training	**Deepfake**	**AIGC**	**IMDL**	**Doc**	**Mean**
Best single-domain	0.8744	0.9717	0.8769	0.7228	0.8615
Unified RGSA	0.9087	0.9468	0.8593	0.7401	0.8637

5.5 Ablation Study

Pretrained backbones. We transfer the same RGSA architecture to other pretrained ViT backbones (CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2023), and SigLIPv2 (Tschannen et al., 2025)) and evaluate their ACC. Tab. 6 shows that RGSA improves over LoRA across all backbones.

Table 6: ACC (%) across pretrained backbones.

Method	DINOv3	CLIP ViT-L/14	DINOv2	SigLIP v2
LoRA	85.4	85.4	82.9	83.6
RGSA	86.4	85.5	84.9	84.9

Role-specialized adaptation paths. We conduct combinatorial ablations over the global pathway, routed experts, and relational anchor, reporting per-domain and mean ACC. The results in Tab. 7 support the stabilizing role of relational anchoring. They also indicate that OFT's constrained orthogonal updates suit the global pathway, whereas LoRA's flexibility is more appropriate for routed residual adaptation. Appendix Tab. 12 further reports the standalone SIL and ACC obtained by activating the global and selective pathways separately.

Table 7: Ablation of relational anchoring and adaptation pathways (ACC).

Anchor	Global	Experts	DF	AIGC	IMDL	Doc	Mean
w/o.	LoRA	LoRA	0.9079	0.9361	0.8581	0.7325	0.8586
w/o.	OFT	LoRA	0.9093	0.9459	0.8585	0.7250	0.8596
with	OFT	OFT	0.8986	0.9426	0.8542	0.7228	0.8546
with	LoRA	OFT	0.9036	0.9429	0.8559	0.7270	0.8574
with	LoRA	LoRA	0.9070	0.9388	0.8583	0.7411	0.8613
with	OFT	LoRA	0.9087	0.9468	0.8593	0.7401	**0.8637**

5.6 Routed Expert Activation Analysis

We analyze stabilized LoRA expert activations in the later stages of training. Tab. 8 measures domain-pair routing differences by averaging total variation (TV) distance across blocks, where 0 indicates identical expert distributions and 1 indicates completely different distributions. Per-block activation patterns are provided in Appendix Fig. 10.

Table 8: Mean total-variation distance between domain-pair expert activations.

Routing	DF–AIGC	DF–IMDL	DF–Doc	AIGC–IMDL	AIGC–Doc	IMDL–Doc	ACC
4 experts + Top-1	0.041	0.035	0.043	0.033	0.046	0.043	0.856
8 experts + Top-1	0.618	0.689	0.777	0.327	0.514	0.617	0.861
8 experts + Top-2	0.538	0.605	0.680	0.264	0.448	0.530	0.864

If domains relied on distinct artifacts, routing TV would approach 1. Four experts show little domain separation, whereas eight experts produce clear separation with residual coupling, supporting both domain-specific and partially shared artifacts. AIGC and IMDL tend to activate similar experts, suggesting that their strong conflict may arise from opposing update demands on similar artifacts. RGSA mitigates interference through more appropriate capacity allocation.

6 Conclusion

In conclusion, this paper reveals and quantifies the mechanism by which heterogeneous artifacts progressively lead to feature-space collapse during joint optimization in FID. Further analysis shows that different artifacts exhibit different cross-domain sharing scopes. Based on this, we propose Role-Specialized Global-Selective Adaptation (RGSA), which uses global and selective pathways to represent different sharing scopes, respectively. Experimental results show that RGSA can effectively mitigate cross-domain interference and feature-space collapse, while outperforming existing state-of-the-art methods. Further analyses validate the sharing-scope-based role specialization hypothesis and provide a mechanism-based guiding principle for unified forensic representation learning.

AI USE STATEMENT

In this work, we used generative AI tools to assist with grammar correction, wording refinement, paraphrasing, and improving the clarity and logical flow of author-written text. We also used them to assist with literature retrieval and understanding of relevant prior work.

All AI-assisted text was reviewed and edited by the authors, and all cited works were independently read and verified against the original sources.

We did not use generative AI tools for research ideation, methodology development, or experimental design. All research concepts, ideas, methodological decisions, experiments, and analyses were developed and conducted by the authors.

We take responsibility for the final content of this work, including all text, claims, analyses, citations, and other artifacts produced or modified with the aid of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. This study involved neither human participants nor animal experiments. All datasets used were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure that the results presented in this paper are reproducible. The experimental setup, including training steps, model configurations, and hardware details, is described in detail in Sections 5.1 and G.1. Additionally, all datasets in this paper are publicly available, ensuring consistent and reproducible evaluation results. We believe these measures will enable other researchers to reproduce our work and further advance the field.

REFERENCES

Afiya Ayman, Ayan Mukhopadhyay, and Aron Laszka. Ensemble prediction of task affinity for efficient multi-task learning. *arXiv preprint arXiv:2602.18591*, 2026.

Ollin Boer Bohan. Megalith-10m. https://huggingface.co/datasets/madebyollin/megalith-10m, June 2024. URL https://huggingface.co/datasets/madebyollin/megalith-10m. Accessed: 2024-10-07.

Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4103–4112. IEEE, 2022.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European conference on computer vision*, pp. 103–120. Springer, 2020.

Jiaxuan Chen, Jieteng Yao, and Li Niu. A single simple patch is all you need for ai-generated image detection, 2024a. URL https://arxiv.org/abs/2402.01123.

Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14165–14173. IEEE, 2021.

Zhongxi Chen, Shen Chen, Taiping Yao, Ke Sun, Shouhong Ding, Xianming Lin, Liujuan Cao, and Rongrong Ji. Enhancing tampered text detection through frequency feature fusion and decomposition. In *European Conference on Computer Vision*, pp. 200–217. Springer, 2024b.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13455–13465. IEEE, 2025.

Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Aayush Dhakal, Subash Khanal, Srikumar Sastry, Jacob Arndt, Philipe Dias, Dalton Lunga, and Nathan Jacobs. Simlbr: Learning to detect fake images by learning to detect real images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35472–35482, 2026.

Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020. URL https://arxiv.org/abs/2006.07397.

Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China summit and international conference on signal and information processing*, pp. 422–426. IEEE, 2013.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Bo Du, Xuekang Zhu, Xiaochen Ma, Chenfan Qu, Kaiwen Feng, Zhe Yang, Chi-Man Pun, Jian liu, and Ji-Zhe Zhou. Forensichub: A unified benchmark & codebase for all-domain fake image detection and localization. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-2167. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5d8a47fc1cb0a143f149ee07321621c9-Paper-Datasets_and_Benchmarks_Track.pdf.

Bo Du, Xiaochen Ma, Xuekang Zhu, Zhe Yang, Chaogun Niu, Chenfan Qu, Mingqi Fang, Zhenming Wang, Jingjing Liu, Jian Liu, et al. Can we build a monolithic model for fake image detection? sica: Semantic-induced constrained adaptation for unified-yet-discriminative artifact feature space reconstruction. *arXiv preprint arXiv:2602.06676*, 2026.

Mingqi Fang, Ziguang Li, Lingyun Yu, Quanwei Yang, Hongtao Xie, and Yongdong Zhang. Forensic-moe: Exploring comprehensive synthetic image detection traces with mixture of experts. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17772–17782. IEEE, 2025.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 27503–27516. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e77910ebb93b511588557806310f78f1-Paper.pdf.

Google AI Blog. Contributing data to deepfake detection, 2019. URL https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html. Accessed 2025-04-25.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 20606–20615. IEEE, 2023.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5037–5047. IEEE, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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. *arXiv preprint arXiv:2106.09685*, 2021a.

Shu Hu, Yuezun Li, and Siwei Lyu. Exposing gan-generated faces using inconsistent corneal specular highlights. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2500–2504. IEEE, 2021b.

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

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 *International Conference on Learning Representations*, volume 2026, pp. 89421–89443, 2026b.

Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. Autosplice: A text-prompt manipulated image dataset for media forensics. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 893–903. IEEE, 2023.

Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, et al. Legion: Learning to ground and explain for synthetic image detection. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18937–18947. IEEE, 2025.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 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.

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5073–5082. IEEE, 2020a.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5000–5009. IEEE, 2020b.

Yiheng Li, Zichang Tan, Guoqing Xu, Zhen Lei, Xu Zhou, and Yang Yang. Towards generalizable ai-generated image detection via image-adaptive prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21262–21272, 2026.

Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3204–3213. IEEE, 2020c.

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

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 6761–6771, 2021a.

Ming-Hui Liu, Harry Cheng, Xin Luo, and Xin-Shun Xu. Suppressing gradient conflict for generalizable deepfake detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9992–10002. Ieee, 2021b.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11966–11976. IEEE, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017a.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017b. URL https://arxiv.org/abs/1608.03983.

Dongliang Luo, Yuliang Liu, Rui Yang, Xianjin Liu, Jishen Zeng, Yu Zhou, and Xiang Bai. Toward real text manipulation detection: New dataset and new solution. *Pattern Recognition*, 157:110828, 2025.

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

Xiaochen Ma, Xuekang Zhu, Lei Su, Bo Du, Zhuohang Jiang, Bingkui Tong, Zeyu Lei, Xinyu Yang, Chi-Man Pun, Jiancheng Lv, et al. Imdl-benco: A comprehensive benchmark and codebase for image manipulation detection & localization. *Advances in Neural Information Processing Systems*, 37:134591–134613, 2024.

Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2307–2311. IEEE, 2019.

Tai D Nguyen, Aref Azizpour, and Matthew C Stamm. Forensic self-descriptions are all you need for zero-shot detection, open-set source attribution, and clustering of ai-generated images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3040–3050. IEEE, 2025.

Adam Novozámskỳ, Babak Mahdian, and Stanislav Saic. Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 71–80. IEEE, 2020.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8245–8257. IEEE, 2025.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Zeju Qiu, Weiyang Liu, Haiwen Feng, Yuxuan Xue, Yao Feng, Zhen Liu, Dan Zhang, Adrian Weller, and Bernhard Schölkopf. Controlling text-to-image diffusion by orthogonal finetuning. *Advances in neural information processing systems*, 36:79320–79362, 2023.

Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards robust tampered text detection in document image: New dataset and new solution. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5937–5946. IEEE, 2023.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10781–10790. IEEE, 2024.

Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Revisiting tampered scene text detection in the era of generative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 694–702, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in neural information processing systems*, 34:8583–8595, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Huiru Shao, Kaizhu Huang, Wei Wang, Xiaowei Huang, and Qiufeng Wang. Progressive supervision for tampering localization in document images. In *International Conference on Neural Information Processing*, pp. 140–151. Springer, 2023.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Jiamu Sun, Zhiyuan Yan, Ke-Yue Zhang, Taiping Yao, and Shouhong Ding. Dfd-hr: Generalizable deepfake detection via hierarchical routing learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13984–13995, 2026.

Ke Sun, Hong Liu, Taiping Yao, Xiaoshuai Sun, Shen Chen, Shouhong Ding, and Rongrong Ji. An information theoretic approach for attention-driven face forgery detection. In *European conference on computer vision*, pp. 111–127. Springer, 2022.

Zhihao Sun, Haipeng Fang, Juan Cao, Xinying Zhao, and Danding Wang. Rethinking image editing detection in the era of generative ai revolution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3538–3547, 2024.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PmLR, 2019.

GuangJian Team. Venus-defakerone: Unified fake image detection & localization. *arXiv preprint arXiv:2605.14091*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2354–2363. IEEE, 2022a.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8692–8701. IEEE, 2020.

Xinyuan Wang, Yingxin Lai, Zhiming Luo, and Zhihui Liu. Ppm-clip: Probabilistic prompt modeling for generalizable ai-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21316–21325, 2026.

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

Yuxin Wang, Hongtao Xie, Mengting Xing, Jing Wang, Shenggao Zhu, and Yongdong Zhang. Detecting tampered scene text in the wild. In *European Conference on Computer Vision*, pp. 215–232. Springer, 2022b.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22388–22398. IEEE, 2023.

Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9535–9544. IEEE, 2019.

Ziyi Xi, Wenmin Huang, Kangkang Wei, Weiqi Luo, and Peijia Zheng. Ai-generated image detection using a cross-attention enhanced dual-stream network. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1463–1470. IEEE, 2023.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 656–665. IEEE, 2022.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection, 2023. URL https://arxiv.org/abs/2307.01426.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024a.

Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems*, 37:29387–29434, 2024b.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33: 5824–5836, 2020.

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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22289–22299. IEEE, 2023.

Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1053–1061. IEEE, 2018.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in neural information processing systems*, 36:77771–77782, 2023.

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

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, 2026. URL https://arxiv.org/abs/2509.20006.

A BENCHMARK

The public forensic datasets involved in constructing OpenMMSec are: (1) Deepfake: FaceForensics++ (Rossler et al., 2019), FaceShifter (Li et al., 2020a), DFD (Google AI Blog, 2019), DFDC (Dolhansky et al., 2020), CelebDF-v2 (Li et al., 2020c), DF40 (Yan et al., 2024b); (2) AIGC: DiffusionForensics (Wang et al., 2023), GenImage (Zhu et al., 2023), Community Forensics (Park & Owens, 2025); (3) IMDL: CASIAv2 (Dong et al., 2013), IMD2020 (Novozámskỳ et al., 2020), tamperCOCO (Kwon et al., 2022), MIML (Qu et al., 2024), Autosplice (Jia et al., 2023), GRE (Sun et al., 2024), OpenSDI (Wang et al., 2025); (4) Doc: Doctamper (Qu et al., 2023), OSTF (Qu et al., 2025), RTM (Luo et al., 2025). Beyond the diverse faking types, OpenMMSec also covers a comprehensive set of real image sources, including ImageNet (Deng et al., 2009), COCO (Lin et al., 2014), UADFV (Li & Lyu, 2018), CelebA (Liu et al., 2018), FaceForensics++ (Rossler et al., 2019), DFD (Google AI Blog, 2019), FFHQ (Karras et al., 2019), CelebDF (Li et al., 2020c), Visual News (Liu et al., 2021a), VFHQ (Xie et al., 2022), LAION (Schuhmann et al., 2022), Megalith-10M (Boer Bohan, 2024), as well as images collected from online sources.

ForensicHub (Du et al., 2025) also introduces a Uni-FID benchmark, termed the IFF-Protocol, primarily to provide a unified cross-domain training and evaluation interface for different models. Its training set combines several fixed datasets from the Deepfake, IMDL, AIGC, and Document domains, with balanced sampling across domains in each epoch. However, the protocol is not systematically constructed with respect to data granularity, artifact diversity, type balance, data sources, or seen–unseen generalization splits, limiting its suitability for comprehensively evaluating unified FID.

B DETAILS OF SUBDOMAINS

Although image forensics fundamentally aims to distinguish real images from manipulated ones, it has long been divided into four subdomains due to differences in targets and manipulation techniques. To the best of our knowledge, research on unified FID is sparse, with existing methods relying either on multi-model ensembles (Huang et al., 2026a) or proprietary, closed-source detection models (Team, 2026).

B.1 DEEPFAKE

Deepfake detection primarily focuses on whether faces in images or videos have been manipulated, and existing methods can be broadly categorized into naive, spatial, and frequency-based detectors (Yan et al., 2023). Deepfake detection designs artifact extraction strategies with facial priors, such as biometric cues (Ciftci et al., 2020), landmarks (Hu et al., 2021b), and facial motion patterns (Haliassos et al., 2021).

B.2 AIGC

AI-generated image detection targets images that are entirely synthesized by generative models. As generative models have rapidly advanced in recent years, they can now produce fake images that are nearly indistinguishable to the human eye. AIGC detection primarily relies on model-specific artifacts, such as reconstruction anomalies in diffusion models (Wang et al., 2023) and frequency irregularities in GANs (Wang et al., 2020), and has attracted substantial attention in recent years.

B.3 IMDL

Image manipulation detection and localization focuses on partial edits in natural images (Zhu et al., 2026) and thus produces two outputs: an image-level prediction of real or fake, and a pixel-level classification of manipulated regions (Ma et al., 2024). IMDL typically relies on hand-crafted artifact extraction strategies, such as noise (Wu et al., 2019; Zhou et al., 2018), frequency (Kwon et al., 2022; Wang et al., 2022a), and edge artifacts (Chen et al., 2021; Ma et al., 2023). In recent years, the rise of generative models has posed new challenges for IMDL, as partial-region manipulations based on generative models have also spurred the development of new detection methods (Wang et al., 2025).

B.4 DOCUMENT

Document image manipulation localization targets text-level manipulations. Such manipulations often manifest as inconsistencies between text and background (Shao et al., 2023); however, compared to other domains, the artifacts are much more subtle, as document images contain far less information than complex faces or natural scenes. Recent methods typically exploit block artifact grids (Qu et al., 2023) or inconsistencies among characters (Wang et al., 2022b).

However, unified FID without distinguishing domains is becoming increasingly important (Du et al., 2025; Huang et al., 2026a). On the one hand, domain-specific approaches create silos that hinder scientific progress; on the other hand, the rapid evolution of generative models is blurring the boundaries between domains. In real-world scenarios, forensic detectors cannot know the domain of a test image in advance, further highlighting the necessity of FID.

C DOMAIN SOTAS INVOLVED IN OUR WORK

C.1 DEEPFAKE

CapsuleNet (Nguyen et al., 2019) applies capsule networks to digital forensics, aiming to detect forged images and videos across diverse scenarios (e.g., replay attacks and CNN-generated fakes), leveraging capsule routing agreement to improve detection on challenging forgeries.

Recce (Cao et al., 2022) is an end-to-end reconstruction–classification learning framework for face forgery detection, where reconstruction learning over genuine faces helps mine common real-face features, and reconstruction discrepancies serve as cues that are jointly optimized with classification to better separate real and fake faces.

Sia (Sun et al., 2022) is a forgery detection approach that introduces a self-information metric into attention, proposing a plug-and-play Self-Information Attention module that emphasizes informative regions and recalibrates feature responses to improve detection performance and generalization.

DFD-HR (Sun et al., 2026) is a generalizable deepfake detection framework that introduces hierarchical routing into visual foundation models, adaptively selecting informative layers and tokens while leveraging a Mixture-of-Experts architecture to improve representation diversity and cross-domain generalization.

C.2 AIGC

UnivFD (Ojha et al., 2023) shows that standard trained real-vs-fake classifiers generalize poorly to new generative models and proposes learning-free fake detection by operating in a pretrained vision–language feature space, using simple nearest-neighbor or linear probing to achieve strong generalization to unseen diffusion and autoregressive models.

Effort (Yan et al., 2024a) identifies an "asymmetry phenomenon" in AI-generated image detection where models overfit limited fake patterns and become low-rank, and proposes SVD-based orthogonal subspace decomposition: freezing principal components to preserve pretrained knowledge while adapting remaining components to learn forgery patterns for better generalization.

AIDE (Yan et al., 2025) combines high-level semantic features from CLIP with low-level features extracted from high- and low-frequency image patches, jointly exploiting semantic and artifact cues for more generalizable AI-generated image detection.

CO-SPY (Cheng et al., 2025) addresses generalization and post-processing robustness (e.g., JPEG) in synthetic image detection by enhancing semantic features (e.g., semantic anomalies) and pixel/artifact features (e.g., low-level differences) separately and then adaptively integrating them.

LEGION (Kang et al., 2025) is an MLLM-based synthetic image analysis framework that jointly performs artifact detection, localization, and textual explanation, supported by the fine-grained SynthScars dataset to provide more interpretable forgery analysis.

SimLBR (Dhakal et al., 2026) reframes AI-generated image detection as learning a tight decision boundary around real images, introducing Latent Blending Regularization in the DINOv3 feature space to treat diverse fakes as a sink class and improve cross-generator generalization.

IAPL (Li et al., 2026) proposes Image-Adaptive Prompt Learning, dynamically adjusting prompts for each test image through conditional information learning and test-time token adaptation to better generalize to unseen generators.

PPM-CLIP (Wang et al., 2026) replaces static classification with probabilistic prompt modeling, generating input-conditioned prompt distributions and incorporating patch-wise contrastive learning to capture novel artifacts and improve cross-domain detection.

C.3 IMDL

IML-ViT (Ma et al., 2023) is a ViT-based benchmark paradigm for image manipulation localization, motivated by the need to capture non-semantic discrepancies via explicit comparisons between manipulated and authentic regions, and designed with high-resolution capacity, multi-scale features, and edge supervision to work with limited data.

Trufor (Guillaro et al., 2023) is a forensic framework for trustworthy image forgery detection and localization that fuses high-level and low-level traces through a transformer-based architecture combining RGB content with a learned noise-sensitive fingerprint.

Mesorch (Zhu et al., 2025) introduces a mesoscopic perspective for manipulation localization by integrating microscopic traces and macroscopic semantic changes, using parallel CNNs (micro details) and Transformers (macro information) with multi-scale orchestration to improve performance, efficiency, and robustness.

RelayFormer (Huang et al., 2026b) is a unified image and video manipulation localization framework that handles variable-resolution inputs without resizing, introducing Global Local Relay tokens to exchange global semantic or temporal context across local sub-images while preserving fine-grained forensic traces for efficient and scalable localization.

C.4 Document

FFDN (Chen et al., 2024b) targets document image tampering detection by jointly modeling spatial and frequency cues: a Visual Enhancement Module makes subtle traces more visible, while a Wavelet-like Frequency Enhancement explicitly decomposes and preserves high-frequency details crucial for detecting weak tampering artifacts.

C.5 Unified FID

SICA (Du et al., 2026) targets unified Fake Image Detection (FID) across Deepfake, AIGC, image manipulation, and document forgery, identifying artifact feature-space collapse as a key limitation of monolithic detectors. It introduces Semantic-Induced Constrained Adaptation to reconstruct a unified-yet-discriminative artifact space, using semantic priors to preserve cross-domain sharing while maintaining subdomain-specific forensic cues.

D Cross-Domain Interaction Analysis

We provide a more comprehensive analysis of the LoRA rank in Tab. 9 and Tab. 10. A more accurate characterization is that an excessively small LoRA rank indeed leads to insufficient representational capacity. However, once the rank reaches a moderate level, the performance saturates: further increasing the rank no longer improves performance or alleviates conflicts.

D.1 A Comprehensive Analysis of LoRA Rank

Table 9: Gradient cosine similarity C across domain pairs for each model.

Model	DF–AIGC	DF–IMDL	DF–Doc	AIGC–IMDL	AIGC–Doc	IMDL–Doc	**Mean**
LoRA ($r = 8$)	0.017	0.024	0.008	-0.052	-0.009	0.011	0.000
LoRA ($r = 16$)	0.016	0.021	0.007	-0.040	-0.007	0.010	0.001
LoRA ($r = 64$)	0.015	0.017	0.005	-0.036	-0.005	0.009	0.001
LoRA ($r = 144$)	0.009	0.008	0.004	-0.023	-0.001	0.006	0.001

Table 10: Relative cross-domain harm R and ACC for each model. In each domain-pair cell, the upper value follows the domain order in the heading and the lower value is the reverse direction.

Model	DF–AIGC	DF–IMDL	DF–Doc	AIGC–IMDL	AIGC–Doc	IMDL–Doc	**Mean**	ACC
LoRA ($r = 8$)	0.384 0.041	-0.034 0.014	0.162 0.076	0.443 0.837	0.228 0.575	0.127 0.020	0.239	0.846
LoRA ($r = 16$)	0.379 -0.002	-0.081 -0.035	0.136 0.048	0.352 0.797	0.188 0.556	0.141 0.014	0.208	0.854
LoRA ($r = 64$)	0.393 0.009	0.028 0.019	0.158 0.082	0.297 0.681	0.173 0.525	0.101 0.041	0.209	0.851
LoRA ($r = 144$)	0.453 0.015	-0.007 0.042	0.156 0.097	0.238 0.685	0.128 0.556	0.103 0.049	0.210	0.855

D.2 Cross-Domain Harm Comparison for RGSA

Fig. 3 and Fig. 4 report the domain-pair matrices of gradient cosine similarity and relative cross-domain harm, respectively, for RGSA, LoRA-Rank16, LoRA+Router LoRA, and OFT+Router LoRA. The results show that RGSA consistently mitigates heterogeneous cross-domain interference across these measures.

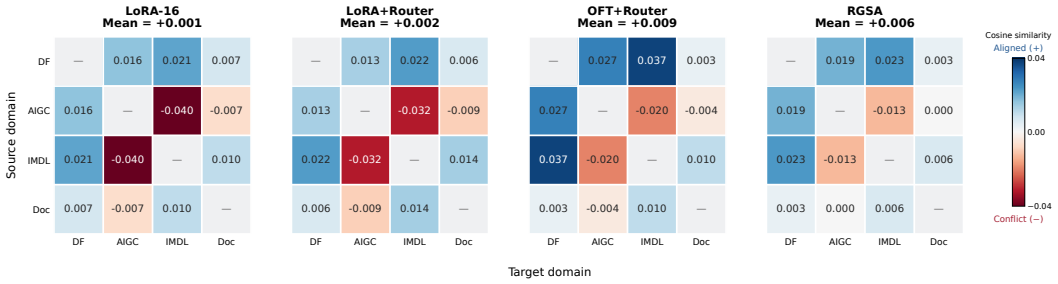

Figure 3: Gradient cosine similarity across domain pairs for RGSA and comparison models.

D.3 Supplementary Evidence for the Long-Term Transfer Metric

In the main text, we quantify ΔM using ACC to characterize long-term transfer in discriminative performance. Since SIL more directly reflects within-domain feature separability and thus feature-space collapse, we additionally compute SIL-based ΔM on the original 1024-dimensional features. Fig. 5 reports the domain-pair matrices of SIL-based ΔM for LoRA-Rank16 and RGSA, together with the ACC-based ΔM of RGSA for comparison.

D.4 Supplementary Definition of Relative Cross-Domain Harm

The original cross-domain harm metric (R) measures the effect of a cross-domain update relative to an in-domain update. Consequently, an update that is beneficial in absolute terms may still be

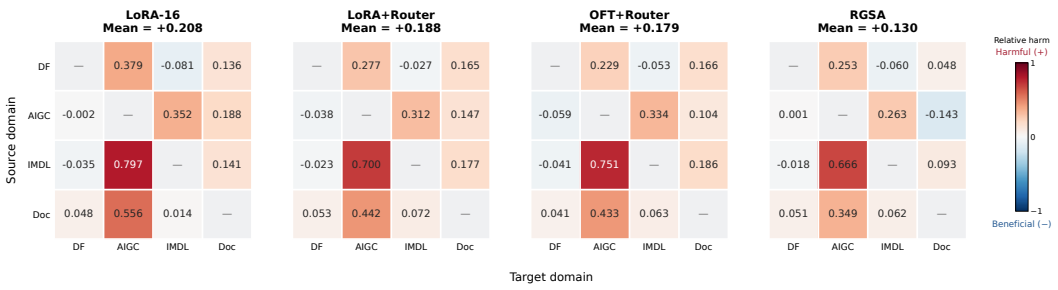

Figure 4: Relative cross-domain harm across directed domain pairs for RGSA and comparison models.

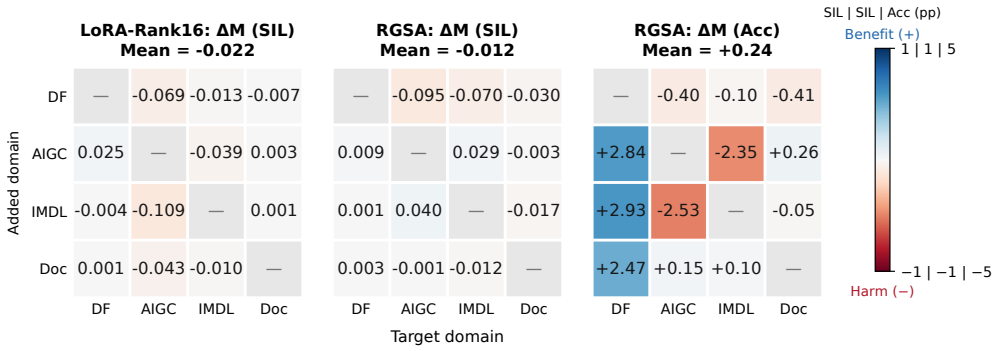

Figure 5: Long-term transfer matrices measured using SIL and ACC.

regarded as relatively harmful if its benefit is weaker than that of an in-domain update, and vice versa. We adopt this relative formulation because in-domain updates are not always beneficial in practice, making the target domain's own learning dynamics a more appropriate reference for cross-domain harm.

$$R^*_{A\to B} = \begin{cases} \frac{\delta_{A\to B}}{|\delta_{B'\to B}|+\epsilon} & \delta_{A\to B} \times \delta_{B'\to B} > 0, \\ \frac{\delta_{A\to B} - \delta_{B'\to B}}{|\delta_{B'\to B}|+\epsilon} & \text{otherwise.} \end{cases}$$

We report the domain-pair matrices of R^* for LoRA-Rank16 and RGSA in Fig. 6. The overall patterns remain consistent with those obtained using R, while R^* generally yields lower harm values. This indicates that some cross-domain updates are beneficial in absolute terms, although their gains are weaker than those produced by in-domain updates. Such observations further support our hypothesis that forensic domains are not completely isolated, but retain partially shareable information. Fig. 8 additionally reports the full distributions of R^* to characterize the actual cross-domain interaction patterns in greater detail.

D.5 Anomalously Low Conflict under Full Fine-Tuning (FFT)

SICA (Du et al., 2026) highlights the difference between full fine-tuning (FFT) and low-rank adaptation, showing that FFT tends to update more strongly along semantic subspaces. We therefore additionally report the cross-domain interaction metrics of DINOv3 under FFT, as shown in Fig. 7. FFT exhibits relatively low cross-domain conflict, yet, consistent with the observations in SICA, it also shows limited generalization on the test set. This contrast suggests that the dominant conflicts in unified FID are more closely associated with heterogeneous artifact representations than with semantic features, and further indicates that effective unified detection critically depends on learning transferable artifact representations.

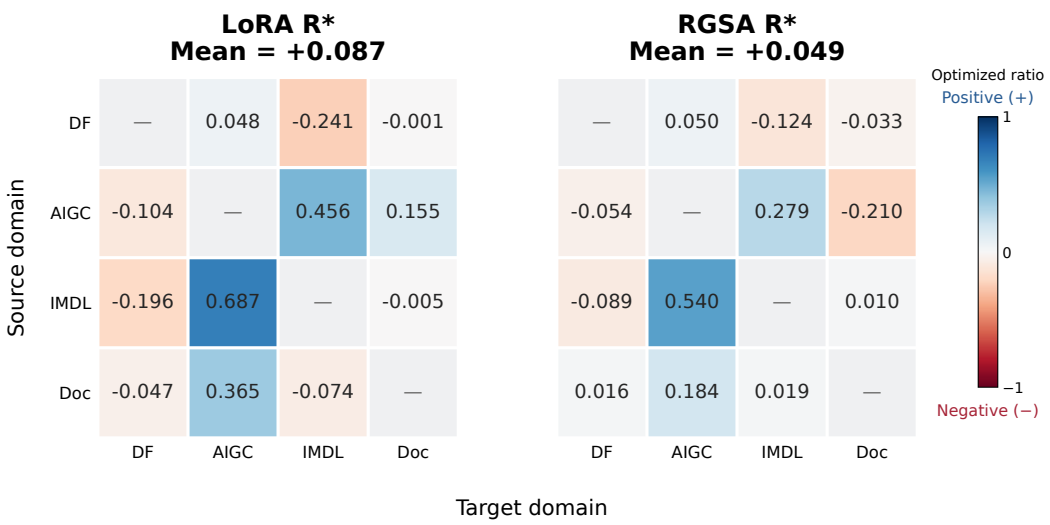

Figure 6: Domain-pair matrices of the alternative relative cross-domain harm measure R^*.

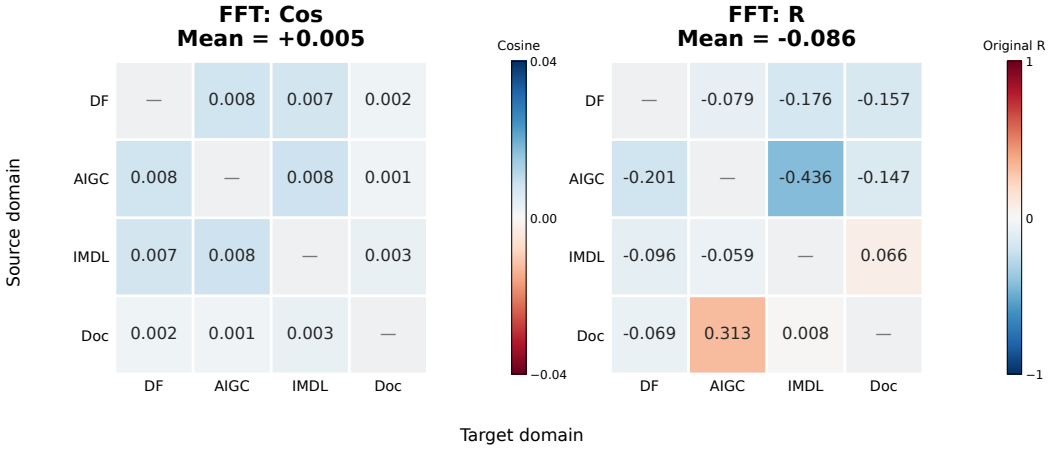

Figure 7: Cross-domain interaction metrics under full fine-tuning.

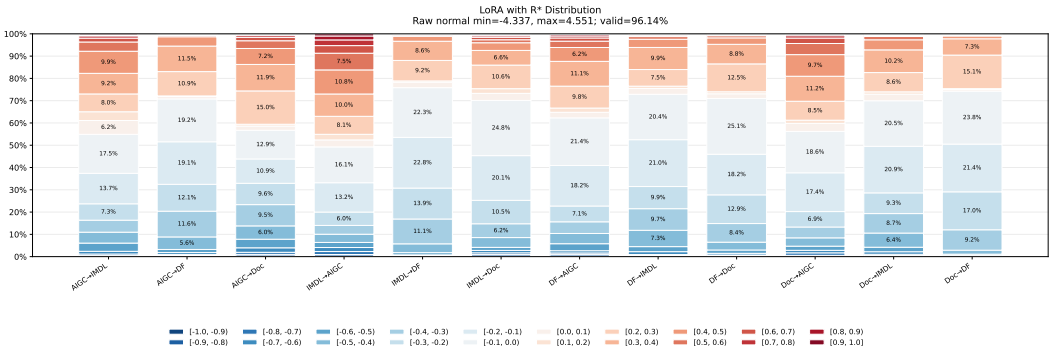

Figure 8: Distributions of the alternative relative cross-domain harm measure R^*.

E VALIDATING FEATURE-SPACE COLLAPSE MITIGATION

Together with the quantitative evidence in Tab. 5, the clear separation in Fig. 9 demonstrates that RGSA successfully mitigates feature-space collapse and constructs a unified-yet-discriminative artifact feature space.

Table 11: Silhouette score (SIL) and accuracy (ACC) under single- and joint-domain training. This single-domain approach combines separately trained models, utilizing only the results from the training domain; Colored subscripts on joint means show the change from single-domain training (red: decrease; green: increase).

	SIL					ACC				
Model	DF	AIGC	IMDL	Doc	Mean	DF	AIGC	IMDL	Doc	Mean
LoRA–Single	0.220	0.360	0.133	0.046	0.190	0.8856	0.9731	0.8741	0.7418	0.8687
LoRA–Joint	0.198	0.216	0.077	0.036	$0.132_{-0.058}$	0.8818	0.9306	0.8559	0.7473	$0.8539_{-0.0148}$
RGSA–Single	0.240	0.374	0.146	0.053	0.203	0.8744	0.9717	0.8769	0.7228	0.8615
RGSA–Joint	0.301	0.298	0.100	0.052	$0.188_{-0.015}$	0.9087	0.9468	0.8593	0.7401	$0.8637_{+0.0022}$

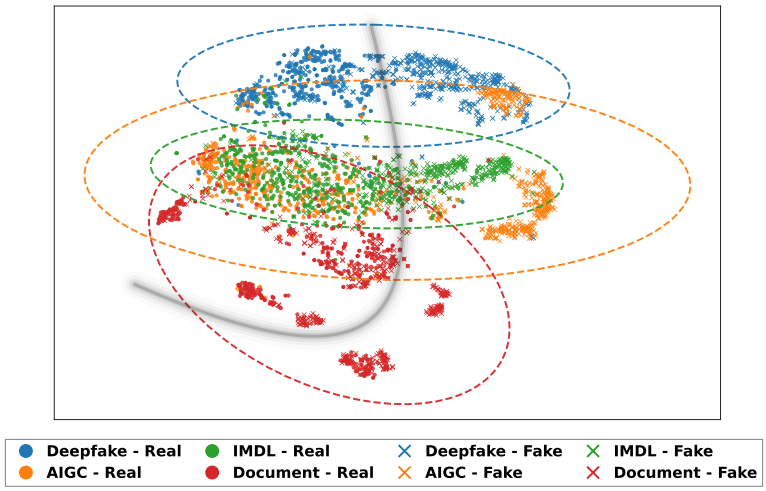

Figure 9: t-SNE visualization of the unified-yet-discriminative artifact feature space learned by RGSA under joint-domain training.

F ROUTING EXPERTS AND GLOBAL–SELECTIVE PATHWAY ANALYSIS

F.1 LORA EXPERT ACTIVATION ANALYSIS

Fig. 10 reports detailed LoRA-expert activation patterns for 4-Experts+Top-1, 8-Experts+Top-1 and 8-Experts+Top-2 including the domain-wise activation of each expert in every DINOv3 block. Statistics are averaged over the stabilized late-training stage (Epochs 7–9). Since the four domains contain different numbers of training samples, raw activation counts are not directly comparable; we therefore normalize each domain to an equal prior of $1/4$. The 4-Experts+Top-1 setting shows little domain differentiation, whereas 8-Experts+Top-2 exhibits clear domain-differentiated routing while retaining partial cross-domain coupling, consistent with the intended role of the expert pathway.

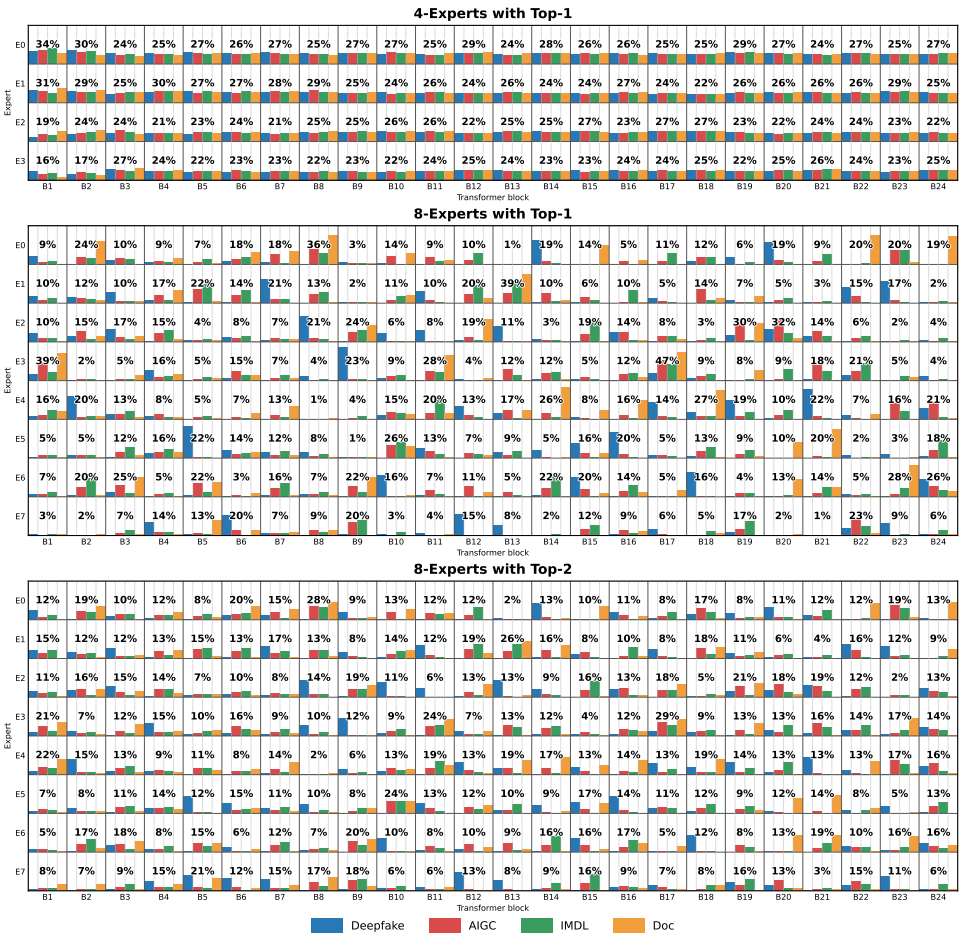

Figure 10: Domain-wise LoRA expert activation patterns across DINOv3 blocks.

F.2 Global and Selective Pathway Analysis

Tab. 12 compares the standalone representation capability of the Global and Selective pathways across Frozen/LoRA, OFT+Router LoRA, Anchor-LoRA (A-LoRA)+Router LoRA, and RGSA (A-OFT+Router LoRA). We also include a comparison with OFT updates without the Global–Selective separation. SIL is computed directly in the original 1024-dimensional feature space.

Table 12: Standalone representation analysis of the Global and Selective pathways. Gray columns show the mean SIL and ACC.

Pathway	Adapter	SIL					ACC				
		Deepfake	AIGC	IMDL	Doc	Mean	Deepfake	AIGC	IMDL	Doc	Mean
Global	Frozen	0.016	0.015	0.001	0.008	0.010	0.5803	0.4954	0.5214	0.4861	0.5208
	OFT	0.018	0.035	0.003	0.008	0.016	0.8181	0.8659	0.7284	0.6379	0.7626
	A-LoRA	0.021	0.030	0.003	0.008	0.016	0.8047	0.8863	0.7483	0.6041	0.7609
	RGSA	0.019	0.038	0.003	0.008	0.017	0.8007	0.8362	0.7184	0.6167	0.7430
Selective	Frozen	0.274	0.287	0.123	0.051	0.183	0.8781	0.9236	0.8589	0.7275	0.8470
	OFT	0.329	0.338	0.128	0.055	0.213	0.9066	0.9447	0.8575	0.7203	0.8573
	A-LoRA	0.302	0.299	0.136	0.060	0.199	0.9018	0.9346	0.8585	0.7411	0.8590
	RGSA	0.301	0.298	0.100	0.052	0.188	0.9047	0.9467	0.8601	0.7416	0.8633
No Global-Selective	OFT	0.216	0.218	0.078	0.036	0.137	0.9049	0.9343	0.8537	0.7258	0.8547

The results show that the Selective pathway substantially outperforms the Global pathway and already approaches the performance of the full model. However, this does not imply that the representations learned by the Global pathway are ineffective. The Global pathway still exhibits stable standalone classification capability, indicating that it indeed captures globally shared artifacts that are useful for fake image detection. Since a single fake image may exhibit multiple forensic traces (Nguyen et al., 2025), while highly discriminative local cues can already provide sufficient evidence for detection (Chai et al., 2020; Chen et al., 2024a), the contributions of different artifact types to the final prediction need not be additive. As a result, once the Selective pathway has already captured sufficiently discriminative evidence, further combining it with Global features does not necessarily yield an approximately additive performance gain.

This interpretation is further supported by the setting without explicit Global–Selective separation. When all artifacts are modeled within a single pathway, the model can still achieve discriminative performance comparable to that of the Selective pathway, suggesting that highly discriminative non-global artifacts remain mixed within this representation. This observation, in turn, provides evidence that RGSA effectively separates globally shared artifacts from non-globally shared ones.

Moreover, different designs of the Global pathway lead to different performance of the Selective pathway, further supporting our central hypothesis. The role of the Global pathway is not only to model shared information, but also to sufficiently absorb globally shared artifacts and prevent them from leaking into the Selective pathway. Otherwise, globally shared and non-shared artifacts may become entangled again, reducing expert separability and interfering with routing decisions.

Therefore, the contributions of the Global and Selective pathways should not be interpreted as simply additive. The two pathways model artifacts at different sharing scopes: the Selective pathway typically provides stronger direct discriminative signals, while the Global pathway isolates globally shared components and thereby provides a cleaner and more separable representation space for non-globally shared artifacts.

F.3 Global and Selective Heatmaps

Fig. 11 shows activation heatmaps for the Global branch, Selective branch, and full model. Both branches exhibit clear and meaningful responses, while the full model consistently focuses on regions closely associated with the underlying forensic artifacts.

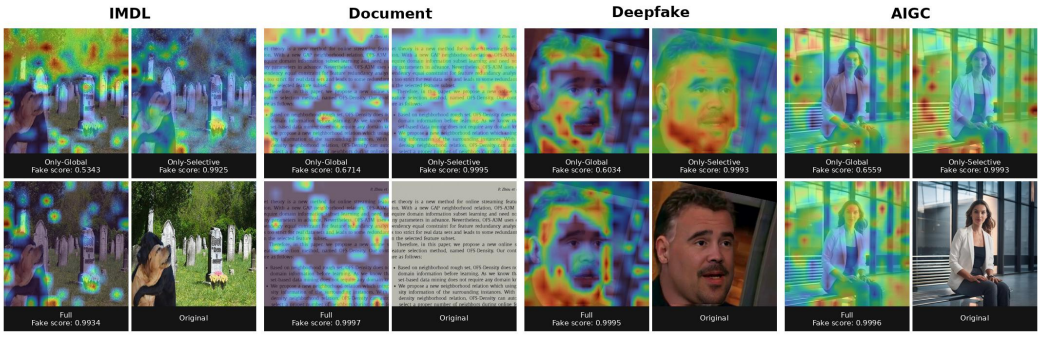

Figure 11: Activation heatmaps of the Global, Selective, and full pathways.

G Additional Experiments

G.1 Implementation Details

We insert Shared OFT and Routed Residual LoRA into the QKV and output projections of the attention module in every ViT block. The model is trained for 10 epochs with a batch size of 64 on four NVIDIA RTX Pro 6000 GPUs, using the AdamW optimizer (Loshchilov & Hutter, 2017a). We use PyTorch (Paszke et al., 2019) for implementation. The learning rate follows a cosine decay (Loshchilov & Hutter, 2017b) schedule from 1×10^{-4} to 1×10^{-5}. Input images

are resized to 224×224 to match the ViT input resolution. The classification objective uses binary cross-entropy (BCE) loss. For the compared detectors, we use the default configurations provided in ForensicHub (Du et al., 2025) and train them on OpenMMSec using the same protocol. All models are trained using the same data augmentation.

G.2 AUC, AP, and F1 Results on OpenMMSec

We report AUC, AP, and F1 in Tab. 13, Tab.14 and Tab. 15, respectively.

Table 13: AUC results on OpenMMSec. The best and second-best domain and overall averages are bold and underlined, respectively; gray columns show domain averages.

Method	**Deepfake**					**AIGC**							**IMDL**			**Doc**			**Overall**
	EFS	FE	FR	FS	**Avg**	GAN	Lat-Diff	Pix-Diff	AR	Comm	Other	**Avg**	Gen	Non-Gen	**Avg**	AIGC	Non-AIGC	**Avg**	**Avg**
ResNet	78.0	58.6	75.3	68.4	70.1	86.0	70.6	71.0	97.7	70.8	79.3	79.2	88.9	75.4	82.2	53.6	86.5	70.1	75.4
EfficientNet	58.7	52.8	59.2	59.2	57.5	73.7	63.2	63.1	84.3	61.8	67.7	69.0	52.5	50.6	51.6	46.7	72.5	59.6	59.4
CapsuleNet	67.7	52.4	72.9	62.2	63.8	82.2	65.3	65.4	90.1	61.9	78.8	74.0	85.7	71.3	78.5	39.6	86.1	62.9	69.8
SegFormer	89.5	80.5	89.8	81.9	85.4	91.8	88.0	84.6	99.8	91.8	77.1	88.9	89.7	78.5	84.1	59.5	87.9	73.7	83.0
Swin	91.6	85.2	86.2	77.3	85.1	93.6	91.5	85.7	99.9	93.6	80.1	90.7	92.3	80.7	86.5	61.4	85.6	73.5	84.0
ConvNeXt	84.7	84.4	90.0	80.3	84.9	92.5	87.0	84.3	99.9	92.2	75.0	88.5	89.5	75.3	82.4	59.5	84.3	71.9	81.9
Recce	68.9	54.1	51.0	55.4	57.4	83.2	71.2	70.2	94.2	67.9	74.4	76.9	59.6	55.2	57.4	62.1	66.5	64.3	64.0
Sia	67.4	53.7	81.0	68.1	67.6	86.3	74.4	74.9	96.7	77.7	81.9	82.0	79.3	55.9	67.6	58.8	89.1	74.0	72.8
IML-ViT	76.0	78.8	85.0	71.1	77.7	87.7	77.3	72.2	99.9	82.3	81.8	83.5	88.8	75.6	82.2	52.7	87.7	70.2	78.4
Trufor	78.6	65.1	83.1	69.8	74.2	88.0	84.8	77.1	99.2	88.8	81.0	86.5	90.7	78.2	84.5	51.8	90.5	71.2	79.1
UnivFD	88.5	94.2	69.8	73.7	81.6	91.5	69.7	78.3	98.8	66.7	87.1	82.0	68.2	84.6	76.4	33.0	79.2	56.1	74.0
FFDN	76.2	51.5	87.2	76.3	72.8	95.9	92.9	94.9	100.0	93.3	74.8	92.0	93.3	80.2	86.8	56.1	84.7	70.4	80.5
Effort	94.7	96.4	87.3	91.3	<u>92.4</u>	95.9	79.3	88.2	99.3	76.8	88.7	88.0	89.0	91.0	**90.0**	48.3	83.6	66.0	84.1
Mesorch	83.0	72.7	87.1	78.8	80.4	88.2	80.1	78.3	99.6	87.5	82.5	86.0	87.6	76.0	81.8	58.9	88.5	73.7	80.5
CO-SPY	80.9	62.3	85.1	71.4	74.9	92.4	88.8	85.5	98.9	94.1	70.2	88.3	84.4	68.0	76.2	52.2	85.7	69.0	77.1
AIDE	93.8	85.4	82.9	74.9	84.2	94.2	86.8	89.1	93.8	86.0	89.0	89.8	89.5	88.2	88.8	48.9	83.6	66.2	82.3
LEGION	93.7	94.7	87.6	86.6	90.6	97.7	87.0	92.6	99.9	92.3	78.3	91.3	90.2	89.2	<u>89.7</u>	58.7	87.8	73.2	86.2
PPMClip	94.5	86.4	86.8	83.4	87.8	98.7	95.9	94.7	99.6	93.5	69.7	92.0	90.1	81.8	85.9	49.6	87.9	68.7	83.6
SIMLBR	92.0	84.1	75.6	69.0	80.2	91.8	85.5	87.1	99.4	83.2	91.0	89.6	67.0	65.8	66.4	50.2	69.2	59.7	74.0
DFD-HR	95.8	68.7	91.0	91.1	86.7	99.7	98.5	97.5	100.0	96.3	56.5	91.4	86.4	79.4	82.9	58.0	87.9	73.0	83.5
RelayFormer	67.9	42.4	81.1	62.9	63.6	98.8	96.7	88.0	100.0	96.2	72.7	92.0	91.8	73.3	82.6	59.0	88.9	74.0	78.0
IAPL	95.5	75.9	84.4	87.0	85.7	97.3	95.4	90.8	99.3	88.4	64.3	89.2	90.7	87.3	89.0	53.3	85.5	69.4	83.3
SICA	95.9	87.1	91.5	92.5	91.8	99.2	98.6	95.5	100.0	97.1	77.3	<u>94.6</u>	90.7	86.4	88.6	62.9	87.6	**75.3**	<u>87.5</u>
RGSA (Ours)	97.9	95.6	93.9	94.3	**95.4**	99.5	98.6	98.2	100.0	97.4	87.3	**96.8**	92.3	85.8	89.1	61.1	88.4	<u>74.8</u>	**89.0**

Table 14: AP results on OpenMMSec. The best and second-best domain and overall averages are bold and underlined, respectively; gray columns show domain averages.

Method	Deepfake					AIGC							IMDL			Doc			Overall
	EFS	FE	FR	FS	Avg	GAN	Lat-Diff	Pix-Diff	AR	Comm	Other	Avg	Gen	Non-Gen	Avg	AIGC	Non-AIGC	Avg	Avg
ResNet	72.4	41.5	67.7	56.9	59.6	65.6	52.5	15.1	84.1	25.8	29.6	45.5	83.6	41.8	62.7	18.3	97.3	57.8	56.4
EfficientNet	46.0	21.2	53.6	52.1	43.2	24.2	43.3	10.4	8.6	17.5	17.2	20.2	39.1	22.9	31.0	15.8	92.8	54.3	37.2
CapsuleNet	61.1	41.0	68.2	52.6	55.7	55.8	48.8	12.2	28.9	19.6	31.7	32.8	81.0	39.0	60.0	14.2	96.6	55.4	51.0
SegFormer	87.0	60.4	88.7	76.5	78.2	75.1	83.2	40.3	95.4	76.8	33.4	67.4	87.4	52.9	70.1	22.4	97.6	60.0	68.9
Swin	89.7	63.8	86.0	70.7	77.5	80.9	87.4	46.1	97.0	79.1	33.4	70.6	90.1	56.9	73.5	23.3	97.1	60.2	70.5
ConvNeXt	83.2	66.9	89.5	74.6	78.5	75.8	82.0	33.1	96.9	79.4	29.7	66.2	87.7	52.4	70.0	23.1	96.8	60.0	68.7
Recce	63.6	37.2	50.5	50.1	50.3	63.5	60.1	17.8	69.6	31.9	35.1	46.3	47.4	27.1	37.3	22.6	93.3	57.9	48.0
Sia	59.7	32.5	77.3	58.7	57.0	63.2	60.6	15.4	79.1	39.6	36.6	49.1	72.0	26.9	49.5	21.8	97.9	59.8	53.9
IML-ViT	68.3	57.4	78.4	59.7	65.9	70.4	63.9	16.9	98.0	50.1	37.5	56.1	84.9	47.9	66.4	17.8	97.6	57.7	61.5
Trufor	72.9	36.3	78.3	58.8	61.6	66.8	77.2	25.9	70.7	69.4	32.8	57.1	86.7	47.9	67.3	17.2	98.0	57.6	60.9
UnivFD	87.7	88.2	68.9	72.1	79.2	75.3	60.3	25.9	83.8	24.3	58.7	54.7	54.8	68.5	61.6	12.9	94.5	53.7	62.3
FFDN	74.0	43.9	85.7	69.2	68.2	86.8	94.1	69.8	99.9	91.4	42.4	80.7	91.2	57.1	74.2	19.8	97.0	58.4	70.4
Effort	94.1	93.3	87.9	91.6	91.7	87.1	67.3	48.4	91.0	30.5	48.0	62.0	85.8	78.7	82.2	16.8	96.2	56.5	73.1
Mesorch	76.1	53.2	85.0	69.3	70.9	68.5	72.3	21.1	92.8	69.6	42.2	61.1	84.1	47.8	65.9	21.6	97.8	59.7	64.4
CO-SPY	79.4	45.8	86.1	65.3	69.2	78.5	85.9	47.1	92.4	81.9	25.6	68.6	81.7	40.6	61.1	19.2	97.2	58.2	64.3
AIDE	93.4	78.5	83.4	72.9	82.1	75.6	74.6	49.7	36.0	45.5	45.2	54.4	85.4	72.8	79.1	18.8	96.4	57.6	68.3
LEGION	93.8	91.0	87.5	86.1	89.6	92.1	80.1	55.2	98.3	73.8	34.0	72.2	86.7	78.9	**82.8**	25.6	97.5	61.6	76.5
PPMClip	93.4	79.9	86.4	82.4	85.5	95.4	93.7	78.8	96.0	78.8	21.7	77.4	88.3	67.5	77.9	20.2	97.6	58.9	74.9
SIMLBR	91.5	74.0	74.9	62.9	75.8	81.3	84.0	66.4	91.1	69.9	80.6	78.9	54.3	40.5	47.4	20.5	93.8	57.2	64.8
DFD-HR	95.2	59.8	92.3	90.7	84.5	98.5	98.0	88.4	99.1	89.9	17.7	81.9	86.8	65.3	76.0	23.9	97.6	60.7	75.8
RelayFormer	62.6	29.0	75.9	52.9	55.1	94.7	95.1	67.1	96.5	86.9	26.0	77.7	89.8	46.0	67.9	21.0	97.9	59.5	65.0
IAPL	94.7	66.3	83.9	86.3	82.8	92.5	92.4	70.1	93.4	67.2	18.0	72.3	89.2	70.3	79.7	23.6	97.0	60.3	73.8
SICA	95.5	82.1	92.4	92.2	90.5	97.0	97.9	77.3	99.7	91.2	34.9	83.0	89.4	70.8	80.1	27.1	97.5	**62.3**	79.0
RGSA (Ours)	97.7	93.1	94.6	94.2	**94.9**	98.1	97.9	90.9	99.8	91.7	65.5	**90.7**	91.1	73.9	82.5	26.1	97.7	61.9	**82.5**

Table 15: F1 results on OpenMMSec. The best and second-best domain and overall averages are bold and underlined, respectively; gray columns show domain averages.

Method	Deepfake					AIGC							IMDL			Doc			Overall
	EFS	FE	FR	FS	Avg	GAN	Lat-Diff	Pix-Diff	AR	Comm	Other	Avg	Gen	Non-Gen	Avg	AIGC	Non-AIGC	Avg	Avg
ResNet	66.5	33.2	68.4	59.1	56.8	47.4	48.2	24.5	20.1	31.5	37.5	34.9	75.0	46.6	60.8	20.9	88.5	54.7	51.8
EfficientNet	56.4	31.6	63.4	60.8	53.1	35.7	47.3	17.7	12.8	27.7	29.2	28.4	44.8	31.5	38.2	24.9	85.7	55.3	43.7
CapsuleNet	57.5	26.5	66.1	54.5	51.1	46.7	39.5	19.7	19.8	22.5	38.1	31.0	69.9	36.6	53.2	12.6	90.7	51.7	46.8
SegFormer	79.4	52.9	81.9	68.4	70.7	61.5	72.0	42.1	35.2	65.2	31.7	51.3	75.0	46.3	60.7	26.8	87.3	57.1	59.9
Swin	80.1	58.1	78.9	63.9	70.2	63.6	78.1	43.5	31.1	65.1	34.6	52.7	78.8	51.0	64.9	27.9	85.2	56.5	61.1
ConvNeXt	72.8	59.9	81.9	62.1	69.1	64.0	68.7	36.5	41.1	68.5	27.7	51.1	75.8	46.0	60.9	27.7	83.9	55.8	59.2
Recce	57.9	29.0	49.1	52.5	47.1	55.8	21.4	14.9	59.5	20.9	27.2	33.3	6.4	3.5	4.9	7.3	61.8	34.5	30.0
Sia	55.4	31.4	73.1	58.6	54.6	49.9	50.7	23.5	24.8	42.5	42.0	38.9	66.0	31.7	48.8	25.7	89.8	57.7	50.0
IML-ViT	63.3	45.7	76.6	64.8	62.6	52.6	53.3	24.0	28.6	47.8	38.6	40.8	72.1	41.9	57.0	11.7	86.0	48.9	52.3
Trufor	66.0	39.3	75.2	55.1	58.9	54.5	68.6	33.0	28.5	58.1	35.6	46.4	74.7	40.8	57.7	18.7	91.9	55.3	54.6
UnivFD	63.5	40.7	65.6	64.9	58.7	58.2	48.8	29.8	26.8	29.6	51.1	40.7	51.5	59.5	55.5	18.9	92.2	55.5	52.6
FFDN	62.5	35.5	79.2	59.0	59.1	77.2	84.3	64.4	58.7	82.3	30.9	66.3	75.4	32.7	54.0	24.0	83.8	53.9	58.3
Effort	84.4	73.6	78.3	83.2	79.9	65.0	54.7	41.7	28.8	33.9	52.0	46.0	74.5	71.0	**72.8**	18.8	86.1	52.4	62.8
Mesorch	72.4	45.5	79.7	70.0	66.9	53.1	60.9	29.5	27.7	56.2	41.4	44.8	72.4	44.7	58.6	26.6	88.6	57.6	57.0
CO-SPY	68.6	34.8	75.4	55.8	58.6	58.6	73.2	38.1	28.2	62.9	23.3	47.4	67.3	32.5	49.9	25.7	86.0	55.9	53.0
AIDE	76.3	51.8	73.6	64.9	66.6	62.0	69.9	43.0	26.8	51.0	53.0	51.0	74.2	66.2	70.2	21.2	88.4	54.8	60.6
LEGION	81.1	67.2	78.2	76.0	75.6	74.0	65.5	52.3	42.3	64.9	32.1	55.2	75.5	69.8	72.7	26.3	86.7	56.5	65.0
PPMClip	84.2	65.1	79.6	71.8	75.2	80.9	85.5	64.3	54.7	70.4	14.2	61.7	76.3	59.4	67.8	20.0	87.6	53.8	64.6
SIMLBR	81.6	65.9	45.9	27.7	55.3	73.2	69.7	61.1	69.2	64.3	72.9	68.4	19.9	24.2	22.0	6.3	57.4	31.9	44.4
DFD-HR	88.9	54.9	84.6	79.6	77.0	90.4	94.0	77.8	69.9	83.7	11.9	71.3	77.0	54.2	65.6	27.2	87.0	57.1	67.7
RelayFormer	60.8	32.1	75.1	47.7	53.9	84.1	87.7	61.5	55.2	78.4	18.0	64.1	77.4	31.5	54.5	28.1	88.6	**58.4**	57.7
IAPL	85.9	58.2	74.6	76.3	73.7	71.1	85.1	52.2	40.6	60.4	12.8	53.7	77.3	59.7	68.5	25.0	83.8	54.4	62.6
SICA	88.9	73.5	84.6	83.9	82.7	87.6	92.4	68.7	62.7	82.8	27.6	70.3	77.7	59.3	68.5	28.8	86.0	57.4	69.7
RGSA (Ours)	92.4	84.6	86.6	85.4	**87.3**	88.2	92.9	78.0	67.5	83.9	60.5	**78.5**	79.2	60.1	69.7	28.5	86.3	57.4	**73.2**

G.3 Experiments on Subdomain Benchmarks

We additionally evaluate RGSA on subdomain benchmarks, using DeepfakeBench (Yan et al., 2023) for Deepfake detection and GenImage (Zhu et al., 2023) for AIGC detection. Tab. 16 and Tab. 17 report the two sets of results, respectively. RGSA achieves the highest within-domain average AUC on DeepfakeBench and the highest average accuracy on GenImage.

Table 16: **AUC on the Deepfake benchmark.** Models are trained on FaceForensics++ (c23) (Rossler et al., 2019) and evaluated on within-domain and cross-domain test sets. Baseline results follow DeepfakeBench (Yan et al., 2023) and SICA. Each Avg is the mean of its test-set columns. The best and second-best group averages are highlighted in bold and underlined, respectively.

Method	Within domain						Cross domain							
	FF-c40	FF-DF	FF-F2F	FF-FS	FF-NT	Avg	CDFv1	CDFv2	DFD	DFDC	DFDCP	FaceShifter	UADFV	Avg
Meso4	0.5920	0.6771	0.6170	0.5946	0.5701	0.6102	0.7358	0.6091	0.5481	0.5560	0.5994	0.5660	0.7150	0.6185
MesoIncep	0.7278	0.8542	0.8087	0.7421	0.6517	0.7569	0.7366	0.6966	0.6069	0.6226	0.7561	0.6438	0.9049	0.7096
CNN-Aug	0.7846	0.9048	0.8788	0.9026	0.7313	0.8404	0.7420	0.7027	0.6464	0.6361	0.6170	0.5985	0.8739	0.6881
Xception	0.8261	0.9799	0.9785	0.9833	0.9385	0.9413	0.7794	0.7365	0.8163	0.7077	0.7374	0.6249	0.9379	0.7629
EfficientB4	0.8150	0.9757	0.9758	0.9797	0.9308	0.9354	0.7909	0.7487	0.8148	0.6955	0.7283	0.6162	0.9472	0.7631
Capsule	0.7040	0.8669	0.8634	0.8734	0.7804	0.8176	0.7909	0.7472	0.6841	0.6465	0.6568	0.6465	0.9078	0.7257
FWA	0.7357	0.9210	0.9000	0.8843	0.8120	0.8506	0.7897	0.6680	0.7403	0.6132	0.6375	0.5551	0.8539	0.6940
X-ray	0.7925	0.9794	0.9872	0.9871	0.9290	0.9350	0.7093	0.6786	0.7655	0.6326	0.6942	0.6553	0.8989	0.7192
FFD	0.8237	0.9803	0.9784	0.9853	0.9306	0.9397	0.7840	0.7435	0.8024	0.7029	0.7426	0.6056	0.9450	0.7609
CORE	0.8194	0.9787	0.9803	0.9823	0.9339	0.9390	0.7798	0.7428	0.8018	0.7049	0.7341	0.6032	0.9412	0.7583
Recce	0.8190	0.9797	0.9779	0.9785	0.9357	0.9382	0.7677	0.7319	0.8119	0.7133	0.7419	0.6095	0.9446	0.7601
UCF	0.8399	0.9883	0.9840	0.9896	0.9441	0.9492	0.7793	0.7527	0.8074	0.7191	0.7594	0.6462	0.9528	0.7738
F3Net	0.8271	0.9793	0.9796	0.9844	0.9354	0.9412	0.7769	0.7352	0.7975	0.7021	0.7354	0.5914	0.9347	0.7533
SPSL	0.8174	0.9781	0.9754	0.9829	0.9299	0.9367	0.8150	0.7650	0.8122	0.7040	0.7408	0.6437	0.9424	0.7747
SRM	0.8114	0.9733	0.9696	0.9740	0.9295	0.9316	0.7926	0.7552	0.8120	0.6995	0.7408	0.6014	0.9427	0.7635
Sia	0.7367	0.9621	0.9492	0.9537	0.9143	0.9032	0.7298	0.7424	0.8102	0.7236	0.7834	0.6608	0.7270	0.7396
Effort	0.8062	0.9928	0.9839	0.9806	0.9625	0.9452	0.8822	0.8772	0.9438	0.8194	0.8711	0.7981	0.9702	<u>0.8803</u>
SICA	0.8344	0.9946	0.9866	0.9855	0.9739	<u>0.9550</u>	0.9067	0.9068	0.9436	0.8241	0.8763	0.8302	0.9740	**0.8945**
RGSA (Ours)	0.8269	0.9973	0.9917	0.9954	0.9815	**0.9586**	0.8624	0.8620	0.9283	0.8346	0.8662	0.8535	0.9333	0.8772

Table 17: **Accuracy (%) on the GenImage AIGC benchmark.** We adopt GenImage (Zhu et al., 2023) as the benchmark, training only on data generated by Stable Diffusion v1.4 (Rombach et al., 2022) and testing on other generators. We use the value reported in GenImage and SICA. The best and second-best results are highlighted in bold and underlined, respectively.

Method	Midjourney	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg
ResNet-50	54.9	99.9	99.7	53.5	61.9	98.2	56.6	52.0	72.1
DeiT-S	55.6	99.9	99.8	49.8	58.1	98.9	56.9	53.5	71.6
Swin-T	62.1	99.9	99.8	49.8	67.6	99.1	62.3	57.6	74.8
CNNSpot	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
F3Net	50.1	99.9	99.9	49.9	50.0	99.9	49.9	49.9	68.7
GramNet	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9
SICA	65.5	100.0	99.9	51.1	66.1	99.7	64.4	55.6	<u>75.3</u>
RGSA (Ours)	67.0	100.0	99.9	54.7	68.4	99.8	72.0	58.0	**77.5**