NOT JUST A MATTER OF SCALE: RECOVERING FORENSIC TRACES TO BUILD A FORENSIC-ENHANCED VISION BACKBONE

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation models expose useful but uneven forensic evidence. We study whether a compact DINOv3-L backbone can recover evidence absent from its CLS readout, using DINOv3-7B only as a training-time teacher. Rank–nullity and orthogonal projection characterize this bottleneck and show when student patch features can reduce reconstruction error. We propose **RELIC**, which uses frozen teacher-relative residuals to guide persistent collectors across student layers; a decoder combines the collected evidence with the projected CLS into one Restored Token. A training-only Local Evidence-Aware Director (LEAD) further guides forensic readout. Rather than matching teacher layers directly, RELIC fixes a residual coordinate system first and then trains collection and readout in stages; a controlled ablation shows that this separation avoids unstable joint transfer. At inference, the teacher and LEAD are removed, leaving a representation usable for standalone classification or plug-and-play detector replacement. On AIGIBench Setting II, RELIC reaches 85.3% domain-macro accuracy, compared with 77.6% for AIDE, 82.3% for DGS-Net, and 75.5% for RA-Det with its original DINOv3-L interface; it reaches 91.4% on Chameleon. Probes and corruption tests further show that the recovered residual carries usable complementary evidence.

1 INTRODUCTION

Advances in image generation have made visual content easier to synthesize while making its origin harder to establish. Early synthetic-image detectors exploited artifacts shared by convolutional generators (Wang et al., 2020); frequency analyses linked some such cues to generator upsampling, and later studies examined spectral and autocorrelation regularities across GAN and diffusion outputs (Frank et al., 2020; Corvi et al., 2023). Diffusion-specific work has also tested reconstruction errors and local intrinsic-dimension features (Tan et al., 2024; Wang et al., 2023; Lorenz et al., 2023). As generator families evolve, however, these cues may not remain discriminative. Pretrained visual encoders provide a stronger foundation: CLIP enables transferable detection with lightweight classifiers (Radford et al., 2021; Ojha et al., 2023), while DINOv2, DINOv3, and masked-image pretraining provide increasingly capable self-supervised representations (Oquab et al., 2024; Siméoni et al., 2025; He et al., 2022).

Recent forensic methods mainly improve how pretrained backbones are exploited. FatFormer introduces image- and frequency-aware adaptation, AIDE combines global semantic and local forensic evidence, Effort preserves pretrained structure through orthogonal adaptation, and B-Free controls semantic correlations through content-aligned data (Liu et al., 2024; Yan et al., 2025a;b; Guillaro et al., 2025). Other approaches study diffusion inversion for unseen generators, training-set diversity, frequency-guided reconstruction, semantic-pixel fusion, and discrepancy cues (Cazenavette et al., 2024; Park & Owens, 2025; Chu et al., 2025; Cheng et al., 2025; Yang et al., 2025). Open-world localization and challenging real-world evaluation have also become explicit test settings (Wang et al., 2025; Li et al., 2025a). These methods improve detector inputs, training diversity, or evaluation coverage; a complementary question is whether a compact backbone already exposes additional forensic evidence through its internal patch features.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

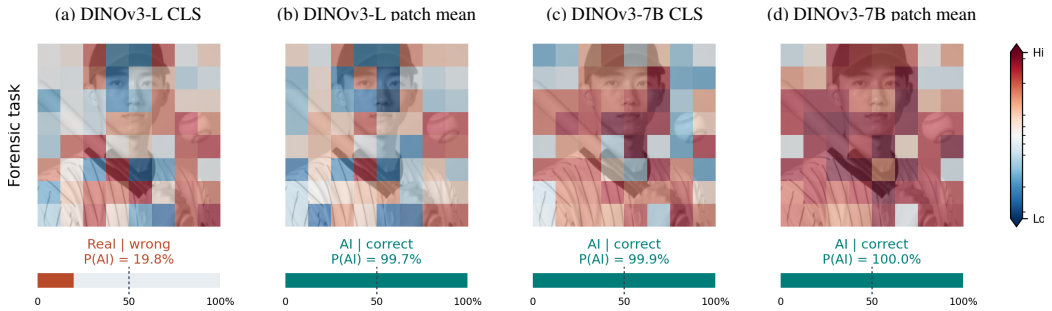

Figure 1: Forensic readouts from four frozen DINOv3 interfaces on the same generated image. The labels above the panels identify the Large CLS, Large patch-mean, 7B CLS, and 7B patch-mean interfaces; each panel shows the local response map and resulting forensic probability.

DINOv3 provides a revealing setting. Its compact variants retain much of the capability of larger models on conventional vision tasks. For example, DINOv3-L achieves 90.2% on ImageNet-ReaL and 54.9 mIoU on ADE20K. The corresponding DINOv3-7B scores are 90.4% and 55.9 mIoU on those two benchmarks (Siméoni et al., 2025).

Image forensics presents a complementary picture. In our evaluation, directly adapting DINOv3 already surpasses most CLIP-based forensic frameworks, while DINOv3-L + LoRA reaches 79.5% domain-macro accuracy on AIGIBench, compared with 86.8% for fine-tuned DINOv3-7B + LoRA using a CLS-only head. This contrast motivates our analysis of how compact global representations expose forensic evidence.

We therefore ask how much useful forensic evidence remains inside the compact student beyond its default global representation. A 1024-dimensional student CLS projected into a 4096-dimensional teacher space cannot create new sample-space directions, while orthogonal projection identifies the teacher component unavailable to this projected CLS. This residual defines the target that the student global feature alone does not recover.

Patch representations provide additional evidence within the same student. Under least-squares reconstruction, adding patch-derived features reduces error exactly by the component of the teacher residual captured by their additional feature span. This motivates recovering complementary information from local student representations rather than merely expanding the global CLS readout.

Based on this insight, we propose **RELIC**, **Residual Evidence Learning via Inter-layer Collection**. RELIC calibrates and freezes student-to-teacher CLS projectors to define stable residual targets. Persistent collector tokens aggregate complementary evidence from student patch representations across multiple depths, and a frozen reconstructor decodes the collected memory into a residual correction. This correction is combined with the projected student CLS to form a single Restored Token. During training, a Local Evidence-Aware Director (LEAD) provides auxiliary forensic supervision. The teacher and LEAD are removed at inference, leaving a compact forensic-enhanced representation that can be used directly for classification. It can also plug into existing detection frameworks.

Our main contributions are summarized as follows:

- **Representation diagnosis:** We identify a task-selective forensic representation gap across DINOv3 model scales and characterize how complementary forensic evidence can be exposed through compact patch representations.
- **Method:** We propose **RELIC**, which uses fixed teacher residual supervision to guide persistent inter-layer collection of complementary evidence from the student's own patch representations and compresses it into a single Restored Token.
- **Plug-and-play backbone:** We expose the recovered representation as a drop-in forensic-enhanced backbone: downstream frameworks retain their decision heads and evaluation rules while replacing only the feature extractor, enabling immediate accuracy gains without retraining the host detector.

2 Related Work

AI-generated image forensics. Cross-generator detectors have used shared synthesis artifacts, frequency cues, diffusion reconstruction, and local intrinsic dimension (Wang et al., 2020; Frank et al., 2020; Corvi et al., 2023; Wang et al., 2023; Lorenz et al., 2023). CLIP-based detectors and recent methods add frequency-aware adaptation, local evidence, inversion, broader generator coverage, semantic-pixel fusion, or discrepancy cues (Ojha et al., 2023; Liu et al., 2024; Yan et al., 2025a; Cazenavette et al., 2024; Park & Owens, 2025; Chu et al., 2025; Cheng et al., 2025; Yang et al., 2025). Open-world benchmarks expose further distribution shifts (Wang et al., 2025; Li et al., 2025a); B-Free, Effort, DGS-Net, and TAP address shortcut control, adaptation, or token aggregation (Guillaro et al., 2025; Yan et al., 2025b; 2026; Abdullah et al., 2026). These methods modify detector inputs, training, or evaluation. RELIC instead asks whether a fixed compact backbone already contains complementary forensic evidence in its patch representations.

Foundation models and representation transfer. DINO, iBOT, masked autoencoders, DINOv2, and DINOv3 progressively strengthen transferable global and local features (Caron et al., 2021; Zhou et al., 2022; He et al., 2022; Oquab et al., 2024; Siméoni et al., 2025). Distillation methods transfer predictions or internal relations (Hinton et al., 2015; Romero et al., 2015; Chung et al., 2020; Zagoruyko & Komodakis, 2017; Park et al., 2019; Tian et al., 2022; Touvron et al., 2021), but output agreement need not preserve every linearly recoverable direction (Nielsen et al., 2026). RELIC freezes the compact student and uses the teacher-relative target to guide cross-layer collection from its patch tokens at multiple depths.

3 Problem Formulation

We study a frozen DINOv3-L student whose default CLS token may omit forensic evidence that remains in its patch features. A prediction head with an m-dimensional bottleneck constrains at most m local directions, leaving a nontrivial null space in the original representation. Thus, matching image-level outputs does not establish that the underlying forensic directions have been transferred.

More precisely, write a prediction head as $h = g \circ b$, where b maps a d-dimensional feature to an m-dimensional bottleneck. Its local Jacobian obeys

$$J_h(z) = J_g(b(z))J_b(z), \qquad \operatorname{rank} J_h(z) \le m, \qquad \dim\ker J_h(z) \ge d - m. \tag{1}$$

The null-space directions are invisible to first-order head matching. Our objective is therefore to measure and recover feature-space evidence directly, rather than treating agreement of a low-dimensional prediction head as evidence of representation equivalence.

3.1 The recoverable residual

At a paired anchor, augment the student CLS matrix with a bias column, $S \in \mathbb{R}^{n\times(d_s+1)}$, and let $T \in \mathbb{R}^{n\times d_t}$ collect teacher CLS features. With $\Pi_S = SS^\dagger$, the best affine CLS readout leaves the orthogonal residual

$$R_\perp = (I-\Pi_S)T, \qquad \min_A \|T-SA\|_F^2 = \|R_\perp\|_F^2, \qquad R_{\mathrm{cal}} = T-SA_0 = R_\perp+(\Pi_S T-SA_0), \tag{2}$$

where A_0 is the calibrated projector used by RELIC. The orthogonal residual $R_\perp$ and the operational target R_{cal} coincide only when $SA_0 = \Pi_S T$; otherwise the operational target also carries a calibration term inside the student span.

Let V be a readout of student patch features and $V_\perp = (I - \Pi_S)V$. The exact least-squares gain from adding these features is

$$(I - \Pi_S)R_{\mathrm{cal}} = R_\perp, \Delta_{\mathrm{patch}} = \|\Pi_{V_\perp}(I - \Pi_S)T\|_F^2. \tag{3}$$

Patch features help only when their residualized span overlaps the orthogonal residual $R_\perp$. RELIC uses the calibrated target R_{cal} for supervision, while this identity characterizes the complementary span analytically. The complete rank argument, projection identities, and proofs are provided in Appendix A and Appendix B.

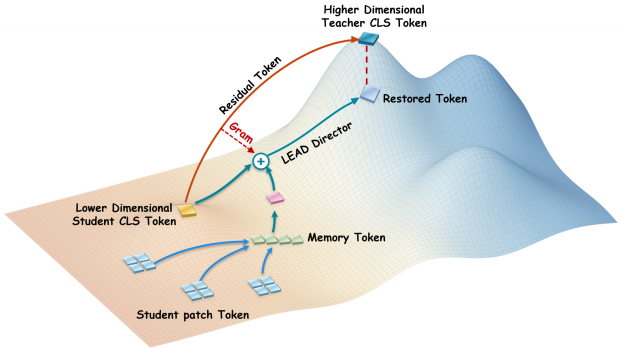

Figure 4: LEAD guidance and downstream integration. The auxiliary director supplies forensic guidance during training; the deployed path retains only the Restored Token interface.

3.2 Why direct transfer is insufficient

A direct layer-wise objective asks the student to agree with teacher features at selected depths, but it does not identify which teacher component is absent from the student's CLS span. In the controlled Setting II ablation, four-key-layer matching reaches 81.10%, a joint transfer objective falls to 54.52%, and staged residual collection reaches 85.26% (Appendix C, Table 16).

4 RELIC: Residual Evidence Learning via Inter-layer Collection

RELIC uses a calibrated teacher-relative residual target with a frozen DINOv3-7B teacher used only during training. Persistent four-slot memory reads patch features at four student depths, and a frozen decoder converts the collected memory into a residual correction. LEAD supplies training-only forensic guidance; inference retains a single Restored Token.

4.1 Residual target and codec

The student is DINOv3-L with forensic LoRA and the teacher is DINOv3-7B; both are frozen during collector training. At blocks $\{6, 12, 16, 24\}$ and $\{10, 20, 27, 40\}$, respectively, affine projectors map the 1024-dimensional student CLS to the 4096-dimensional teacher space. For anchor l, calibration minimizes a SmoothL1-plus-cosine distance and then freezes

$$P_l(s) = W_l\,\mathrm{LN}(s) + \beta_l, \qquad r_l = \mathrm{sg}(t_l - P_l(s_l))\,, \tag{4}$$

$$\ell(u, v) = \mathrm{SmoothL1}(u, v) + 1 - \cos(u, v). \tag{5}$$

The residual r_l is tokenized and reconstructed with a shared codec,

$$T_l = E(r_l), \qquad \widetilde{r}_l = D(T_l), \qquad \mathcal{L}_{\mathrm{rec}} = \sum_{l=1}^{4} \ell(\widetilde{r}_l, r_l), \tag{6}$$

which fixes the coordinate system for the four collector slots. The tokenizer is training-only; the decoder D remains in the deployed path. Full calibration equations and optimizer settings are in Appendix A.

4.2 Persistent collection and restored readout

Four persistent memory slots are updated at each anchor by gated cross-attention over the current patch tokens, with shared updater weights and depth/slot embeddings. Excluding CLS and storage tokens from the keys and values prevents the collector from simply copying the global shortcut.

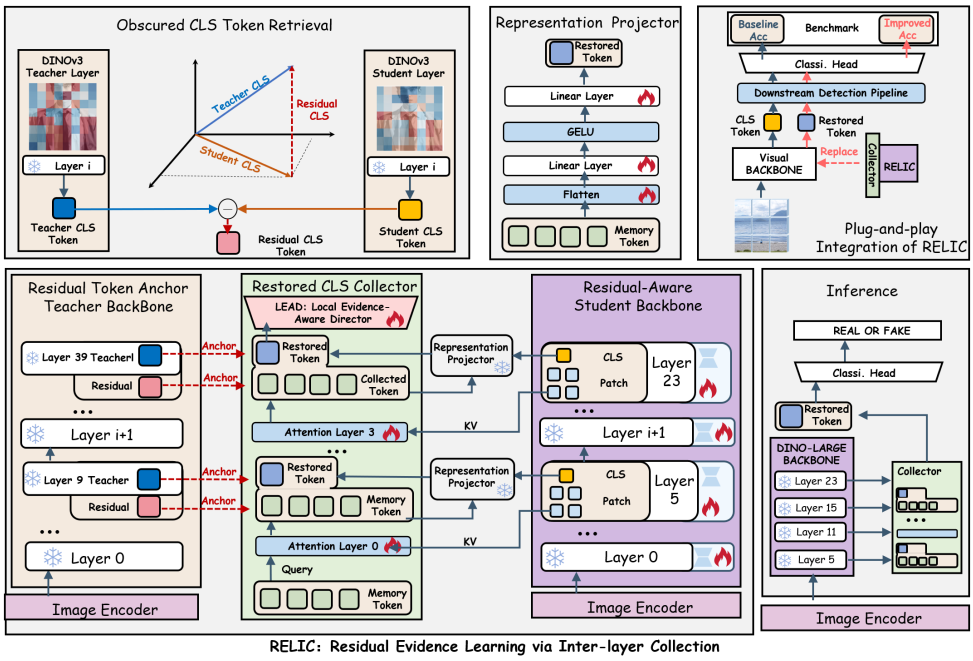

Figure 2: Overview of RELIC. A frozen teacher defines residual supervision relative to the projected student CLS, while persistent collector tokens gather evidence from student patch representations and reconstruct a correction for the final Restored Token.

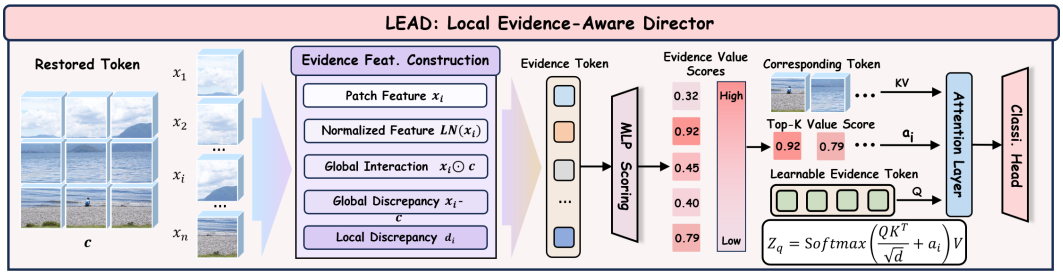

Figure 3: Persistent collector tokens progressively read student patch representations across depth and compress the accumulated evidence into a compact memory for residual reconstruction.

With M_0 reset for each image, one update is

$$Q_l = M_{l-1} + e^{\text{slot}} + e_l^{\text{depth}}, \tag{7}$$

$$\Delta_l = \text{MHA}(\text{LN}_q(Q_l), \text{LN}_x(X_l), \text{LN}_x(X_l)), \tag{8}$$

$$G_l = \sigma(\text{MLP}_g([Q_l; \Delta_l])), \tag{9}$$

$$M_l = Q_l + G_l \odot \Delta_l + \text{FFN}(\text{LN}(Q_l + G_l \odot \Delta_l)). \tag{10}$$

The recurrent memory accumulates complementary evidence across depth. At anchor l, the restored representation and deployed classifier are

$$\hat{t}_l = P_l(s_l) + D(M_l), \qquad z_s = W_c \, \text{LN}(\hat{t}_4) + b_c.$$

Thus the deployed classifier sees one global Restored Token even though the evidence is collected from multiple spatial representations. The collector details and complete implementation equations are given in Appendix A.

4.3 LEAD AND TRAINING OBJECTIVE

LEAD scores final-layer patches using the restored token and forms complementary sparse and query-based readouts. For final-layer patch X_i, a task-trained score is computed as

$$a_i = \mathrm{MLP}_{\mathrm{score}}\left([X_i; \mathrm{LN}(X_i); X_i - c; X_i \odot c; d_i]\right), \qquad c = A(\hat{t}_4), \tag{11}$$

where d_i is a local-grid smoothness statistic. The sparse branch selects the highest-scoring quarter of patches, while a four-query branch adds a_i as an attention bias. Their joint auxiliary readout is trained with

$$z_a = \mathrm{MLP}_{\mathrm{cls}}([c; z_f; z_q; |c - z_f|; c \odot z_f]), \qquad \mathcal{L}_a = \mathrm{CE}(z_a, y). \tag{12}$$

LEAD is discarded after training. The collector objective combines classification, slot reconstruction, residual/restored-feature alignment, and attention diversity; all loss definitions and weights are listed in Appendix A. The teacher, tokenizer, LEAD, and auxiliary readouts are absent at inference.

4.4 DEPLOYMENT

At Stage 2, only the collector, restored-feature adapter, LEAD head, and global classifier are optimized; the student, teacher, projectors, tokenizer, and decoder are frozen. The deployed path keeps the DINOv3-L student with forensic LoRA, the persistent collector, the final projector, the decoder, and a normalized classifier. The teacher, tokenizer, first three projectors, LEAD, and auxiliary readouts are removed. A lightweight mapping exposes the Restored Token to downstream detectors that require the native 1024-dimensional interface. The full objective is

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_a + 0.1\mathcal{L}_{\mathrm{anchor}} + 0.01\mathcal{L}_{\mathrm{div}}, \qquad \mathcal{L}_s = \mathrm{CE}(z_s, y), \tag{13}$$

with anchor weights $(0.5, 0.75, 1, 2)$. This staged separation is the reason direct layer-wise transfer is evaluated as a distinct control rather than used as the main training objective.

5 EXPERIMENTS

We organize the evaluation around four questions: does RELIC improve the main benchmark and an in-the-wild set, which representation carries the gain, can the representation replace features in existing detectors, and which components and perturbations explain its behavior? Unless stated otherwise, all AIGIBench numbers use the complete Setting II evaluation protocol and fixed checkpoints from the independent server run.

5.1 EXPERIMENTAL SETUP

All AIGIBench experiments use the standard Setting II protocol, including the 144K ProGAN+SD-v1.4 training protocol and the complete 212,802-image evaluation set. We additionally evaluate the 26,013-image Chameleon collection. The standalone detector uses fixed-logit argmax, while plug-and-play rows use each host detector's evaluator. The DINOv3-7B comparator is task-adapted with LoRA and a CLS-only head; the teacher and LEAD are never used at inference. Dataset, preprocessing, and optimization details are in Appendix A.

5.2 MAIN RESULTS ON AIGIBENCH AND CHAMELEON

AIGIBench. Table 1 compares RELIC with the retained published and host-detector references over the benchmark domains. RELIC reaches 85.3% domain-macro accuracy, improving over AIDE (77.6%) by 7.7 points, DGS-Net (82.3%) by 3.0 points, and RA-Det with its original DINOv3-L interface (75.5%) by 9.8 points. The advantage is broad but generator-dependent: RELIC is especially strong on CommunityAI, SocialRF, and DALLE-3, while the source-wise appendix tables also expose domains where the compact baseline remains competitive. The complete domain table and operating-point measurements are given in Appendix C.

Table 1: AIGIBench Setting II domain-accuracy excerpt (%). The table keeps the horizontal layout of the complete benchmark table while retaining AIDE, DGS-Net, RA-Det, RELIC, and the DINOv3-7B + LoRA capacity reference. The three blocks follow the benchmark domain order; no method-family rows are inserted. The RA-Det row uses the original DINOv3-L feature interface, while the corresponding RELIC plug-in result is reported in Table 8. All rows use the complete 212,802-image evaluation set.

Method	Venue	ProGAN	R3GAN	StyleGAN3	StyleGAN-XL	StyleSwim	WFIR	BlendFace	E4S	FaceSwap
AIDE	ICLR'25	97.2	92.9	88.1	88.7	83.7	71.4	51.5	44.3	52.1
DGS-Net	ICML'26	99.7	90.5	90.7	89.6	89.8	92.3	68.1	56.9	76.4
RA-Det	ICML'26	100.0	94.7	72.6	97.0	91.3	83.2	51.0	52.5	51.0
RELIC (0.36B)	–	96.8	91.8	92.2	93.5	92.5	83.0	58.9	70.0	61.3
DINOv3-7B + LoRA (6.72B)	–	99.7	99.5	94.7	99.5	97.7	99.4	49.7	49.7	50.3

Method	Venue	InSwap	SimSwap	FLUX1-dev	Mid-journey	GLIDE	DALLE-3	Imagen3	SD3	SDXL
AIDE	ICLR'25	50.9	54.9	88.0	76.4	93.4	55.1	89.8	94.3	93.5
DGS-Net	ICML'26	78.7	83.8	88.9	73.0	92.5	41.5	89.9	91.8	92.0
RA-Det	ICML'26	50.6	50.3	79.7	73.6	90.6	73.8	56.1	82.2	92.0
RELIC (0.36B)	–	62.7	62.7	91.9	88.8	88.1	93.8	80.1	95.2	95.5
DINOv3-7B + LoRA (6.72B)	–	50.1	49.7	91.8	93.5	99.2	98.4	75.2	99.6	99.8

Method	Venue	BLIP	Infinite-ID	InstantID	IP-Adapter	Photo-Maker	SocialRF	Community-AI	Mean
AIDE	ICLR'25	96.4	92.2	91.8	90.0	91.7	57.8	54.1	77.6
DGS-Net	ICML'26	98.5	90.7	89.5	90.2	87.9	59.3	54.6	82.3
RA-Det	ICML'26	96.1	86.0	93.7	80.0	73.5	64.7	50.2	75.5
RELIC (0.36B)	–	95.5	93.5	94.1	88.4	76.2	91.0	94.0	**85.3**
DINOv3-7B + LoRA (6.72B)	–	99.6	99.6	99.6	96.0	84.4	96.6	95.9	86.8

AIDE and DGS-Net are published baselines. RA-Det denotes the host detector with its original DINOv3-L interface, while RELIC denotes the standalone 0.36B representation. The gray DINOv3-7B + LoRA row is a fine-tuned CLS-only capacity reference, not the frozen teacher used by RELIC. The complete plug-and-play RA-Det comparison, including the RELIC replacement, is given in Table 8.

Table 2: In-the-wild results on Chameleon (26,013 images). The RELIC row uses the historical Large Epoch-2 Simple deployment branch.

Method	Venue	Acc. (%)	AUC	EER (%)	AP
CNNSpot	CVPR 2020	58.38	0.63	40.62	0.58
FreDect	ICML 2020	63.04	0.75	31.43	0.69
UnivFD	CVPR 2023	65.98	0.80	26.70	0.75
AIDE	ICLR 2025	62.56	0.80	27.47	0.73
Effort	ICML 2025	64.59	0.84	24.28	0.77
TriDetect	AAAI 2026	67.88	0.89	18.43	0.87
DINOv3-L + LoRA	–	80.04	0.89	18.40	0.89
RELIC (ours)	–	**91.44**	**0.96**	**8.59**	**0.96**

This table is an in-the-wild cross-dataset measurement. EER is the mean of FPR and FNR at the ROC point minimizing their absolute difference. Balanced accuracy and class recalls are recorded in the experiment log.

Chameleon. Table 2 evaluates the same deployed interface on an in-the-wild collection. RELIC obtains 91.44% accuracy, 0.96 AUC, 0.96 AP, and 8.59% EER, compared with 80.04%, 0.89, 0.89, and 18.40% for DINOv3-L + LoRA. This corresponds to +11.40 accuracy points, +0.07 AUC, +0.07 AP, and a 9.81-point EER reduction.

5.3 Representation evidence and transfer

The main comparison does not by itself identify which part of RELIC carries the improvement. Equal-capacity probes in Appendix C, Table 12 separate the projected CLS, decoded residual, Restored Token, and a CLS-plus-memory control. On AIGIBench, the residual probe reaches 84.57% accuracy versus 82.43% for projected CLS, while the Restored Token reaches 85.26%, the best ac-

curacy in the probe table. Chameleon shows a different ordering, so the table supports accessible complementary evidence rather than a claim that every feature combination is uniformly beneficial. The full representation-readout table is Appendix C, Table 17.

Native readouts provide a second control for this interpretation. Appendix C, Table 6 evaluates frozen CLS and mean-patch features on semantic, prompt-class, and forensic tasks under the same fixed readout schedule. Their task-dependent ordering shows why the proposed target is forensic rather than a generic scale or transfer target. The complete domain table is given in Appendix C, Table 9; its generator-wise view in Table 10 shows the largest RELIC gains on Midjourney (+16.96 points) and ADM (+8.60), while SD1.4, SD1.5, and Wukong are already near saturation. Grouping the same scores by exact-version overlap (Table 13) yields 89.55% versus 83.57% on the no-exact-version-overlap group, separating transfer to novel generators from performance on the training generator version.

The collector diagnostics in Appendix C, Table 14 support the same mechanism-level reading. The four anchors have low pairwise attention overlap and small top-10% intersection-over-union, while their effective patch counts remain broad. The slots therefore gather related but nonidentical spatial evidence instead of acting as four copies of one pooled patch summary.

5.4 Plug-and-play transfer across detectors

We replace only the feature extractor and keep each host detector's head and evaluation rule unchanged. Under the 25-domain host summaries, the RA-Det interface rises from 75.5% to 80.0%, ForensicConcept rises from 78.9% to 83.3%, and FGTS rises from 50.0% to 55.0% with four collector tokens. Figure 9 shows why these averages should not be read as uniform gains: in the complete 25-domain tables, ForensicConcept improves on 15 domains and regresses on 10, while RA-Det improves on 16 and regresses on 9. The largest changes and the exact threshold conventions are reported in Appendix C, Tables 7 and 8.

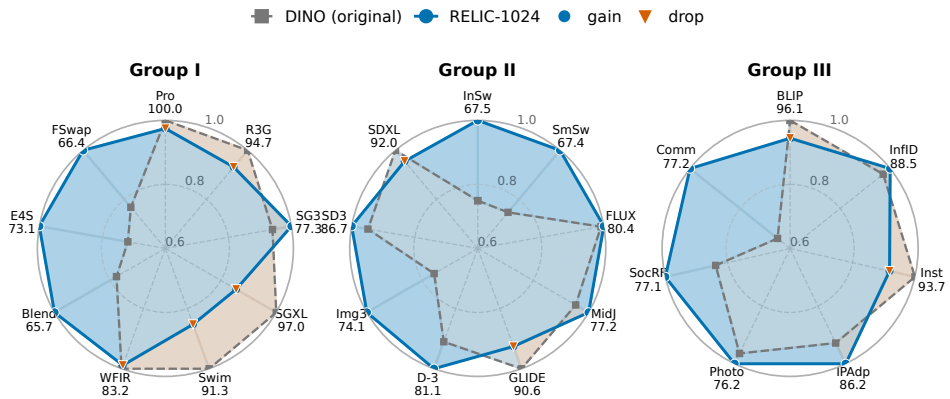

Figure 9: RA-Det per-domain optimal-threshold accuracy on all 25 AIGIBench domains. Each spoke is normalized by the larger accuracy for that domain; blue circles mark gains and orange triangles mark drops for the RELIC-1024 feature replacement. The raw scores are reported in Appendix C, Tables 7 and 8.

5.5 Ablation of the recovery path

Table 3 is a strict component ablation under the fixed Setting II protocol: each row adds the named component to the preceding configuration. LoRA adaptation raises accuracy from 75.32% to 79.17%; adding the collected token raises it to 83.73%, residual correction to 84.19%, and the full Restored Token configuration to 85.26%. The separate direct-distillation study in Appendix C, Table 16, explains the training choice: four-layer matching reaches 81.10%, joint transfer falls to 54.52%, and staged residual collection reaches 85.26%. The same ablation also shows why the staging matters: removing only the residual term from the joint objective recovers 81.96%, which still

trails the staged result. The comparison motivates the order of operations in RELIC: first calibrate the residual coordinates and codec, then learn collection and readout with those coordinates fixed.

Table 3: RELIC component ablation on AIGIBench. Each row adds the named component to the preceding configuration under the fixed Setting II protocol; checkmarks indicate active components or training signals, and LEAD is training-only.

Component	LoRA	Representation projector	Collected token	Residual decoder	LEAD aux. CE	Acc. (%)
DINOv3-L (no fine-tuning)						75.32
DINOv3-L + LoRA	✓					79.17
Collected Token	✓	✓	✓			83.73
Residual correction	✓	✓	✓	✓		84.19
Restored Token (RELIC)	✓	✓	✓	✓	✓	**85.26**

5.6 Robustness and operating points

We apply identical JPEG, blur, resize, and noise transformations to fixed RELIC and DINOv3-L + LoRA checkpoints. RELIC retains a clear margin under severe corruption: at JPEG quality 60 it reaches 78.23% versus 69.29%, and at noise $\sigma = 0.05$ it reaches 81.07% versus 71.80%. Under resizing, both models remain close to their clean scores, while blur produces a larger drop for RELIC at $\sigma = 2$ (81.28% versus 77.65%). The operating-point table further shows higher AIGIBench TPR at 1% FPR (44.28% versus 29.60%) and lower Brier/ECE values (0.1328/12.49 versus 0.1929/18.97), indicating both ranking and calibration improvements. Full paired rows and metric definitions are in Appendix C, Tables 15 and 11.

Under the fixed RTX 4090 protocol in Appendix A, Table 4 reports 2.493 ms per image and 1.430 GiB for RELIC, versus 2.272 ms and 1.212 GiB for DINOv3-L + LoRA. This modest overhead produces a 4096-dimensional forensic-enhanced token for a standalone head or a downstream detector interface.

Taken together, these controls separate recovery from scale: task-specific readouts, generator splits, collector statistics, and direct-transfer ablations all require a stable residual coordinate system before local evidence can be collected. Appendix C, Figure 5, illustrates readout variation on matched images rather than dataset-level frequencies. The domain tables 7 and 8 therefore support a broad but nonuniform plug-and-play gain.

6 Conclusion

RELIC recovers a fixed teacher residual from student patch features and compresses it into a single forensic-enhanced token. Rank–nullity and projection explain the representation gap, while persistent collection and LEAD expose complementary evidence without retaining the teacher at inference. The resulting detector reaches 85.3% domain-macro accuracy and transfers through a 1024-dimensional interface; domain-level and corruption analyses characterize its limits.During training, a DINOv3-7B teacher defines a stable residual target, while a persistent four-slot collector aggregates complementary patch evidence across student depth. LEAD provides label-aware guidance during training but is discarded at inference, leaving a compact single-token pathway. This design exposes richer patch-level forensic evidence while preserving compatibility with existing detection systems.During training, a DINOv3-7B teacher defines a stable residual target, while a persistent four-slot collector aggregates complementary patch evidence across student depth. LEAD provides label-aware guidance during training but is discarded at inference, leaving a compact single-token pathway. This design exposes richer patch-level forensic evidence while preserving compatibility with existing detection systems.

Practically, RELIC is a drop-in forensic-enhanced backbone: it preserves a single-token contract and improves downstream detectors without changing their heads, at modest inference overhead.

AI USE AND REPRODUCIBILITY

Generative AI tools were used to assist with manuscript organization, language editing, and LaTeX formatting. All technical claims, mathematical derivations, experimental results, and conclusions were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide detailed model architectures, optimization settings, training schedules, and computational costs in Appendix A; complete derivations and proofs in Appendix B; and full domain-wise results, robustness evaluations, and controlled ablations in Appendix C. All reported main results use fixed checkpoints without AIGIBench- or Chameleon-based checkpoint selection.

REFERENCES

Ahmed Abdullah, Nikolas Ebert, and Oliver Wasenmüller. TAP into the Patch Tokens: Leveraging Vision Foundation Model Features for AI-Generated Image Detection. arXiv preprint arXiv:2604.26772, 2026. URL https://arxiv.org/abs/2604.26772.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. arXiv preprint arXiv:2104.14294, 2021. URL https://arxiv.org/abs/2104.14294.

George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10759–10769, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Cazenavette_FakeInversion_Learning_to_Detect_Images_from_Unseen_Text-to-Image_Models_by_CVPR_2024_paper.html.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13455–13465, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Cheng_CO-SPY_Combining_Semantic_and_Pixel_Features_to_Detect_Synthetic_Images_CVPR_2025_paper.html.

Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12830–12839, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chu_FIRE_Robust_Detection_of_Diffusion-Generated_Images_via_Frequency-Guided_Reconstruction_Error_CVPR_2025_paper.html.

Inseop Chung, Seonguk Park, Jangho Kim, and Nojun Kwak. Feature-map-level online adversarial knowledge distillation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2006–2015, 2020. URL https://proceedings.mlr.press/v119/chung20a.html.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: From generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 973–982, 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/WMF/html/Corvi_Intriguing_Properties_of_Synthetic_Images_From_Generative_Adversarial_Networks_to_CVPRW_2023_paper.html.

Joel Frank, Thorsten Eisenhofer, Lea Sch"onherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine*

Learning Research, pp. 3247–3258, 2020. URL https://proceedings.mlr.press/v119/frank20a.html.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4 edition, 2013. doi: 10.56021/9781421407944. URL https://www.press.jhu.edu/books/title/10678/matrix-computations.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18685–18694, June 2025. URL https://arxiv.org/abs/2412.17671.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. arXiv preprint arXiv:1503.02531, 2015. URL https://arxiv.org/abs/1503.02531.

Zhenglin Huang, Jason Li, Haiquan Wen, Tianxiao Li, Xi Yang, Lu Qi, Bei Peng, Xiaowei Huang, Ming-Hsuan Yang, and Guangliang Cheng. Rethinking cross-generator image forgery detection through DINOv3, 2025. URL https://arxiv.org/abs/2511.22471.

Chunxiao Li, Xiaoxiao Wang, Meiling Li, Boming Miao, Peng Sun, Yunjian Zhang, Xiangyang Ji, and Yao Zhu. Bridging the gap between ideal and real-world evaluation: Benchmarking ai-generated image detection in challenging scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20379–20389, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Li_Bridging_the_Gap_Between_Ideal_and_Real-world_Evaluation_Benchmarking_AI-Generated_ICCV_2025_paper.html.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is Artificial Intelligence Generated Image Detection a Solved Problem? arXiv preprint arXiv:2505.12335, 2025b. URL https://arxiv.org/abs/2505.12335.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10770–10780, June 2024. URL https://arxiv.org/abs/2312.16649.

Peter Lorenz, Ricard L. Durall, and Janis Keuper. Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 448–459, 2023. URL https://openaccess.thecvf.com/content/ICCV2023W/DFAD/html/Lorenz_Detecting_Images_Generated_by_Deep_Diffusion_Models_Using_Their_Local_ICCVW_2023_paper.html.

Beatrix M. G. Nielsen, Emanuele Marconato, Luigi Gresele, Andrea Dittadi, and Simon Buchholz. Logit Distance Bounds Representational Similarity. arXiv preprint arXiv:2602.15438, 2026. URL https://arxiv.org/abs/2602.15438.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489, June 2023. URL https://arxiv.org/abs/2302.10174.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut,

Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. arXiv preprint arXiv:2304.07193, 2024. URL https://arxiv.org/abs/2304.07193.

Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8245–8257, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Park_Community_Forensics_Using_Thousands_of_Generators_to_Train_Fake_Image_CVPR_2025_paper.html.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for Thin Deep Nets. arXiv preprint arXiv:1412.6550, 2015. URL https://arxiv.org/abs/1412.6550.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. arXiv preprint arXiv:2508.10104, 2025. URL https://arxiv.org/abs/2508.10104.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28130–28139, June 2024. URL https://arxiv.org/abs/2312.10461.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Representation Distillation. arXiv preprint arXiv:1910.10699, 2022. URL https://arxiv.org/abs/1910.10699.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357, 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. URL https://arxiv.org/abs/1912.11035.

Xinchang Wang, Yunhao Chen, Yuechen Zhang, Congcong Bian, Zihao Guo, Xingjun Ma, and Hui Li. RA-Det: Towards Universal Detection of AI-Generated Images via Robustness Asymmetry. arXiv preprint arXiv:2603.01544, 2026. URL https://arxiv.org/abs/2603.01544.

Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. Opensdi: Spotting diffusion-generated images in the open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4291–4301, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_OpenSDI_Spotting_Diffusion-Generated_Images_in_the_Open_World_CVPR_2025_paper.html.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22445–22455, October 2023. URL https://arxiv.org/abs/2303.09295.

Jiazhen Yan, Ziqiang Li, Fan Wang, Boyu Wang, Ziwen He, and Zhangjie Fu. Dgs-net: Distillation-guided gradient surgery for clip fine-tuning in ai-generated image detection. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://github.com/HorizonTEL/DGS-Net. ICML 2026 Spotlight; arXiv:2511.13108.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A Sanity Check for AI-generated Image Detection. arXiv preprint arXiv:2406.19435, 2025a. URL https://arxiv.org/abs/2406.19435.

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. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70268–70288. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/yan25b.html.

Yongqi Yang, Zhihao Qian, Ye Zhu, Olga Russakovsky, and Yu Wu. D^3: Scaling up deepfake detection by learning from discrepancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23850–23859, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_D3_Scaling_Up_Deepfake_Detection_by_Learning_from_Discrepancy_CVPR_2025_paper.html.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sks9_ajex.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer. arXiv preprint arXiv:2111.07832, 2022. URL https://arxiv.org/abs/2111.07832.

Menyanshu Zhou, Ziyin Zhou, Ke Sun, Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, and Rongrong Ji. ForensicConcept: Transferable forensic concepts for AIGI detection. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2606.07034.

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. arXiv preprint arXiv:2306.08571, 2023. URL https://arxiv.org/abs/2306.08571.

A Implementation Details

A.1 Backbones and task adaptation

The student is DINOv3 ViT-L/16, with 24 transformer blocks, width 1024, and 16 attention heads. The teacher is the original DINOv3 ViT-7B/16, with 40 blocks and width 4096. The detector input has 196 spatial patches. Four storage tokens are excluded from the collector's keys and values, and the CLS token is treated separately. Anchor tokens are passed through the backbone's shared final LayerNorm before projection or collection.

LoRA inserts $W_{\text{eff}} = W_0 + (\alpha/r)BA$ into the QKV and output projections of all student attention blocks, with $r = 8$ and $\alpha = 16$. Original pretrained weights remain frozen. Main-model adaptation uses batch size 8 and up to eight epochs, with a head-only first epoch followed by LoRA and local-head updates. The head learning rate is 5×10^{-5}, and the LoRA rate is one quarter of that value. Weight decay is 0.02, cosine scheduling ends at 10^{-6}, and the seed is 42. In addition to classification, this adaptation uses squared cosine regularization between the pooled local feature and CLS with coefficient 0.05, mean evidence-score regularization with coefficient 0.002, and pooling-entropy regularization with coefficient 0.0005. These terms are not added to the subsequent collector-training objective.

A.2 Main-model evaluation protocol

The reported AIGIBench checkpoints come from an independent server run under the fixed Setting II protocol. The main model uses the fixed adaptation epoch-five and collector epoch-two schedule.

A.3 Projector and codec calibration

Each of the four projectors contains LayerNorm followed by a linear map from 1024 to 4096 dimensions. Projectors are independently parameterized. The tokenizer first normalizes the 4096-dimensional residual, applies a 4096–4096–4096 MLP with GELU, reshapes the result into four slots, adds learned slot embeddings, and normalizes each slot. The decoder flattens four 1024-dimensional slots and applies a 4096–4096–4096 MLP with GELU. Encoder and decoder weights are shared across anchors.

Projector calibration and codec calibration each run for one epoch with batch size 2, learning rate 10^{-4}, AdamW weight decay 0.02, and gradient clipping at norm 5. Calibration uses the same Setting II training data as adaptation. The tokenizer is needed only to construct slot targets during training; the decoder remains part of inference. Freezing the codec keeps slot coordinates consistent between calibration and collection. Neither calibration nor the equal number of input and output scalars implies exact reconstruction.

A.4 Collector and LEAD architecture

Collector memory, slot embeddings, and depth embeddings use Gaussian initialization with standard deviation 0.02. The cross-attention block has 16 heads of width 64 and zero attention dropout. Its gate is a 2048–1024–1024 MLP with GELU and a final sigmoid. The feed-forward network is a 1024–4096–1024 MLP with GELU. One updater is shared across all four anchors, and the collector block uses no stochastic-depth layer.

The LEAD adapter contains LayerNorm at width 4096, a linear projection to 1024, GELU, a 1024-dimensional linear map, and a final LayerNorm. The evidence-scoring MLP maps 4097 input coordinates through hidden widths 768 and 384 to a scalar, using GELU and dropout 0.25 after the first hidden layer. The sparse readout selects 49 of the 196 patches. Its complementary attention readout uses four queries and 16 heads, with evidence bias coefficient 1. The auxiliary classifier maps the 5120-dimensional concatenation in Equation 35 through LayerNorm, a 768-dimensional hidden layer, GELU, dropout 0.25, and a two-logit output. Semantic projection subtraction, context mixing, and edge priors are disabled.

During collector training, trainable parameters belong to the collector, restored-feature adapter, LEAD head, and global classifier. A frozen decoder still transmits derivatives to its input; evaluating it under a no-gradient context would incorrectly remove the classification-to-collector path. The teacher and student operate in evaluation mode throughout this stage.

A.5 Loss reduction and computational cost

SmoothL1 uses transition parameter 1 with elementwise mean reduction, whereas cosine terms are averaged over the batch and, for slot supervision, over slots. Let A_{lhki} be collector attention at anchor l, head h, slot k, and patch i. Define $\bar{a}_{lk}$ by averaging heads and normalizing the resulting patch-weight vector to unit Euclidean norm. The diversity term is

$$\mathcal{L}_{\mathrm{div}} = \frac{1}{4}\sum_{l=1}^{4} \frac{1}{BK(K-1)} \sum_{b=1}^{B} \sum_{i \neq j} \left(\bar{a}_{lbi}^{\top} \bar{a}_{lbj}\right)^2, \qquad K = 4. \tag{14}$$

Anchor weights are summed without division by their total. Equation 39 gives the loss coefficients used for the main configuration.

Each collector update includes patch projections with cost $O(Nd_s^2)$ and attention with cost $O(KNd_s)$; the decoder also contributes dense projection cost. The attention map has only $K \times N$ entries per head, but that does not describe the complete module's runtime.

A.6 FORMAL PROBLEM FORMULATION

We seek a compact representation for real-versus-synthetic classification using a frozen DINOv3-L student after task adaptation and a frozen DINOv3-7B teacher during training. The student provides a 1024-dimensional CLS token and spatial patch tokens; the teacher provides a 4096-dimensional CLS token. The analysis below applies rank–nullity, singular value decomposition (SVD), and least-squares projection (Golub & Loan, 2013) to these interfaces. All matrix identities hold for arbitrary finite feature matrices, including rank-deficient ones.

For an image x with forensic label y, write the student CLS and patch features at anchor l as $s_l(x)$ and $X_l(x)$, and write the paired teacher CLS feature as $t_l(x)$. The deployed representation is a single token $z_R(x)$ computed from the student features and passed to a classifier q_ϕ; the teacher is used only to define training-time evidence targets. The formal recovery problem is therefore to construct z_R so that it preserves the teacher-relative forensic component that is not accessible from an affine readout of s_l, while keeping the output width fixed and the inference path independent of the teacher.

This formulation separates three quantities that are otherwise easy to conflate: the orthogonal residual outside the student CLS span, the calibrated teacher-relative residual used as a stable target, and the additional component captured by patch features. The propositions below characterize these quantities for finite feature matrices without assuming that the student or teacher features are full rank. RELIC then solves the corresponding recovery problem by calibrating the residual coordinates first and learning the patch collector and readout in a separate stage.

A.6.1 WHAT PREDICTION-HEAD ALIGNMENT CONSTRAINS

DINOv3 transfers image-level and patch-level predictions through separate heads (Siméoni et al., 2025). If a head is written as $h = g \circ b$, where the bottleneck b has m output coordinates and its input has dimension d, then

$$J_h(z) = J_g(b(z))J_b(z), \qquad \operatorname{rank} J_h(z) \le m, \qquad \dim\ker J_h(z) \ge d - m. \tag{15}$$

Therefore, matching a head leaves first-order directions in its null space unconstrained. This motivates measuring forensic evidence directly rather than inferring it from prediction-head agreement; the complete rank argument is given in Appendix B.

A.6.2 COMPLEMENTARY EVIDENCE BEYOND THE CLS READOUT

At the final anchor, stack affine student CLS features into $S \in \mathbb{R}^{n\times(d_s+1)}$ and teacher CLS features into $T \in \mathbb{R}^{n\times d_t}$. With $\Pi_S = SS^\dagger$, Proposition 1 characterizes the exact least-squares limit of every affine CLS readout.

Proposition 1 (Exact limit of an affine CLS readout)**.**

$$\|T - SA\|_F^2 = \|(I - \Pi_S)T\|_F^2 + \|\Pi_S T - SA\|_F^2, \qquad \min_A \|T - SA\|_F^2 = \|(I - \Pi_S)T\|_F^2, \tag{16}$$

with $\operatorname{rank}(SA) \le d_s + 1$.

The irreducible term is the orthogonal residual $R_\perp = (I - \Pi_S)T$ outside the affine CLS span. For the calibrated projector, $B = SA_0$ and $R_{\text{cal}} = T - B$ obey

$$R_{\text{cal}} = R_\perp + (\Pi_S T - B), \qquad (I - \Pi_S)R_{\text{cal}} = R_\perp. \tag{17}$$

The term $\Pi_S T - B$ lies in the student CLS span, so $R_{\text{cal}} = R_\perp$ only when the calibrated projector attains the least-squares affine fit. RELIC uses R_{cal} as its training target; the patch proposition below concerns $R_\perp$, not the full calibrated target.

A.6.3 WHAT PATCH FEATURES CAN CONTRIBUTE

Let V contain a fixed readout of the student patch features and let $V_\perp = (I - \Pi_S)V$. The additional least-squares gain is exactly

Proposition 2 (Exact contribution of additional features)**.**

$$\min_A \|T - SA\|_F^2 - \min_{A,C} \|T - SA - VC\|_F^2 = \|\Pi_{V_\perp}(I - \Pi_S)T\|_F^2. \tag{18}$$

Thus patch features help precisely when their residualized span overlaps $R_\perp$. RELIC realizes this complement with a learned four-slot collector across depth; proofs of both propositions are included in Appendix B. A controlled Setting II ablation shows why direct layer-wise transfer is insufficient: four-key-layer matching reaches 81.10%, while a fully joint transfer objective falls to 54.52%, compared with 85.26% for the staged residual-collection schedule (Appendix C, Table 16).

A.7 Detailed RELIC method

RELIC enhances the forensic representation of a compact student by recovering complementary evidence from its own patch tokens. During training, a frozen teacher defines a stable residual target relative to the student's global representation, while a persistent collector progressively gathers local evidence across student depth. A training-only Local Evidence-Aware Director (LEAD) further guides this recovery toward real-versus-synthetic discrimination. At inference, the teacher and all auxiliary supervision are removed, leaving a single Restored Token for downstream classification.

A.7.1 Fixed Teacher-Relative Residual Targets

For an image x with label $y \in \{0, 1\}$, let the frozen student produce a CLS token $s_l \in \mathbb{R}^{d_s}$ and patch tokens $X_l \in \mathbb{R}^{N \times d_s}$ at anchor l, while the frozen teacher produces a paired CLS token $t_l \in \mathbb{R}^{d_t}$. Our main model uses $d_s = 1024$, $d_t = 4096$, $N = 196$, and four approximately depth-aligned anchor pairs: student blocks $\{6, 12, 16, 24\}$ and teacher blocks $\{10, 20, 27, 40\}$. Both encoders receive the same augmented image. The teacher is the original self-supervised DINOv3-7B encoder, while the student is DINOv3-L equipped with previously trained forensic LoRA weights. Both are fixed during RELIC training.

Student-to-teacher projection. Before training the collector, we independently calibrate an affine projector at each anchor,

$$P_l(s) = W_l \,\mathrm{LN}(s) + \beta_l, \tag{19}$$

by minimizing

$$\ell(u, v) = \mathrm{SmoothL1}(u, v) + 1 - \cos(u, v), \qquad \mathcal{L}_{\mathrm{proj}} = \sum_{l=1}^{4} \ell(P_l(s_l), t_l). \tag{20}$$

After calibration, all projectors are frozen. We define

$$b_l = P_l(s_l), \qquad r_l = \mathrm{sg}(t_l - b_l), \tag{21}$$

where sg denotes stop-gradient.

The calibrated residual r_l is the fixed teacher-relative target used in training. It equals the orthogonal component outside the student CLS span only when the calibrated projector is the least-squares affine fit; otherwise it also contains a calibration term inside the student span.

Residual tokenization and reconstruction. To provide structured supervision for the collector, a shared Residual Tokenizer E converts each teacher-relative residual $r_l \in \mathbb{R}^{4096}$ into four 1024-dimensional target tokens,

$$T_l = E(r_l) \in \mathbb{R}^{4 \times 1024}. \tag{22}$$

A shared Dimensional Reconstructor D maps these tokens back to the teacher feature space and returns

$$\tilde{r}_l = D(T_l) \in \mathbb{R}^{4096}. \tag{23}$$

We jointly calibrate E and D by requiring the reconstructed residual $\tilde{r}_l$ to match its original target:

$$\mathcal{L}_{\mathrm{rec}} = \sum_{l=1}^{4} \ell(\tilde{r}_l, r_l). \tag{24}$$

After calibration, both modules are frozen. The resulting token representation provides a consistent supervision space for the four collector memory slots. During inference, the tokenizer E is removed, while the frozen reconstructor D maps the collected memory directly into the residual correction.

A.7.2 PERSISTENT INTER-LAYER EVIDENCE COLLECTION

The residual target specifies what should be recovered, but the recovery itself is performed entirely from student representations. RELIC maintains four persistent memory tokens

$$M_l \in \mathbb{R}^{4\times 1024} \tag{25}$$

that evolve across the four student anchors. Each image starts from the same learned initialization M_0, and the memory is reset between images.

At anchor l, the persistent memory is updated by first forming depth-aware slot queries and then reading the current patch representations:

$$\begin{aligned} Q_l &= M_{l-1} + e^{\text{slot}} + e_l^{\text{depth}}, \\ \Delta_l &= \text{MHA}(\text{LN}_q(Q_l), \text{LN}_x(X_l), \text{LN}_x(X_l)), \\ G_l &= \sigma(\text{MLP}_g([Q_l; \Delta_l])). \end{aligned} \tag{26}$$

The attended evidence is written into memory through an element-wise gate, followed by a residual feed-forward update:

$$\tilde{M}_l = Q_l + G_l \odot \Delta_l, \qquad M_l = \tilde{M}_l + \text{FFN}(\text{LN}(\tilde{M}_l)). \tag{27}$$

The same update block is shared across all four anchors. Here, S denotes learned slot embeddings and e_l denotes depth embeddings. Cross-attention uses 16 heads, and the feed-forward network expands the hidden width by a factor of four.

The memory tokens act as persistent queries: each stage reads all patch tokens at the current depth, while the recurrent state allows later stages to refine evidence accumulated from earlier layers. The element-wise gate G_l controls how strongly newly attended evidence is written into memory.

The collector directly reads normalized student patch representations, with CLS and storage tokens excluded from its keys and values. The orthogonalized patch representation introduced in Sec. 3 serves as an analytical reference for complementary feature span, while RELIC realizes this complement through learned collection.

At each anchor, the collected memory is decoded into a residual correction,

$$\hat{r}_l = D(M_l), \tag{28}$$

and combined with the projected student CLS:

$$\hat{t}_l = b_l + \hat{r}_l = P_l(s_l) + D(M_l). \tag{29}$$

The final representation $\hat{t}_4 \in \mathbb{R}^{4096}$ is the ***Restored Token***. The main detector operates only on this single global representation:

$$z_s = W_c\,\text{LN}(\hat{t}_4) + b_c, \qquad \mathcal{L}_s = \text{CE}(z_s, y). \tag{30}$$

Thus, although RELIC collects evidence from multiple local representations, its deployed interface remains a single global token.

A.7.3 LEAD: TRAINING-ONLY FORENSIC GUIDANCE

To prioritize recovered directions that support real-versus-synthetic discrimination, we introduce a *Local Evidence-Aware Director* (LEAD), which supplies label-dependent auxiliary supervision during training.

An adapter maps the final Restored Token to the student feature width,

$$c = A(\hat{t}_4). \tag{31}$$

For final-layer patch X_i, let

$$R_i = \text{LN}(X_i), \tag{32}$$

and let d_i denote its mean Euclidean distance to valid four-connected neighbors on the spatial patch grid. LEAD predicts a scalar evidence score

$$a_i = \text{MLP}_{\text{score}}([X_i; R_i; X_i - c; X_i \odot c; d_i]). \tag{33}$$

The scores are learned directly from image-level labels; patch-level artifact annotations are not used during training.

LEAD forms two complementary local readouts. First, it selects the highest-scoring quarter of patches and applies normalized score weighting:

$$z_f = \sum_{i \in \mathcal{K}} w_i R_i, \qquad w_i = \frac{\exp(a_i)}{\sum_{j \in \mathcal{K}} \exp(a_j)}, \tag{34}$$

where $\mathcal{K}$ denotes the selected patch set.

Second, four learned queries perform score-biased cross-attention over all patches, with a_i added to the corresponding attention logits. Averaging and normalizing the resulting query outputs gives z_q.

The auxiliary prediction is

$$z_a = \mathrm{MLP}_{\mathrm{cls}}\left([c; z_f; z_q; |c - z_f|; c \odot z_f]\right), \qquad \mathcal{L}_a = \mathrm{CE}(z_a, y). \tag{35}$$

The LEAD queries are independent of the persistent collector memory and are discarded after training. LEAD injects task-dependent gradients into the recovery pathway through the Restored Token.

The top-selection branch is piecewise differentiable with respect to a fixed selected set, while the score-biased query branch provides a fully differentiable path from the auxiliary objective to the collector memory. Because the backbone, projector, and decoder parameters remain frozen, LEAD changes how evidence is collected without changing the underlying student representation or residual coordinate system.

A.7.4 Training Objective and Deployment

During collector training, supervision is applied at every anchor. We encourage memory slots to match their tokenizer-defined targets,

$$\ell^l_{\mathrm{slot}} = \mathrm{SmoothL1}(\mathrm{LN}_0(M_l), \mathrm{LN}_0(T_l)) + 1 - \frac{1}{4}\sum_{k=1}^{4} \cos(M_{lk}, T_{lk}), \tag{36}$$

where LN_0 denotes normalization without learned affine parameters.

We additionally supervise the decoded correction and restored representation:

$$\mathcal{L}_{\mathrm{anchor}} = \sum_{l=1}^{4} \omega_l \left[\ell^l_{\mathrm{slot}} + \ell(D(M_l), r_l) + \ell(\hat{t}_l, t_l)\right], \qquad \omega = (0.5, 0.75, 1, 2). \tag{37}$$

The residual identity is

$$\hat{t}_l - t_l = D(M_l) - r_l, \tag{38}$$

the SmoothL1 components of the residual- and restored-feature terms share the same reconstruction error. Their cosine components impose complementary directional constraints: one aligns the predicted correction with the residual target, whereas the other aligns the resulting Restored Token with the teacher representation. We use both constraints in the reported training objective.

To encourage different memory slots to attend to complementary spatial evidence, we further use an attention-diversity penalty $\mathcal{L}_{\mathrm{div}}$, defined as the average squared pairwise cosine similarity between slot attention maps after averaging over attention heads.

The complete Stage-2 objective is

$$\mathcal{L} = \mathcal{L}_s + \mathcal{L}_a + 0.1\mathcal{L}_{\mathrm{anchor}} + 0.01\mathcal{L}_{\mathrm{div}}. \tag{39}$$

During this stage, both backbones, the forensic LoRA parameters, all projectors, the Residual Tokenizer, and the Dimensional Reconstructor are frozen. Only the persistent collector, the main classifier, and the training-only auxiliary path are optimized.

Inference. The deployed model retains only the DINOv3-L student, its frozen forensic LoRA, the persistent collector, the final student-to-teacher projector P_4, the frozen reconstructor D, and the normalized linear classifier. The DINOv3-7B teacher, Residual Tokenizer E, the first three projectors, LEAD, and all auxiliary readouts are removed.

For downstream detectors requiring the native 1024-dimensional student interface, we additionally evaluate a lightweight mapping from the final Restored Token to the required feature width in a dedicated transfer configuration.

The deployed prediction is therefore based on a single Restored Token rather than an ensemble of auxiliary paths. Appendix B gives the classifier stability bound and the complete implementation details for the training-only guidance path.

A.8 Deployment details

The deployed encoder keeps the student, collector, final projector, decoder, and normalized classifier; the teacher, tokenizer, auxiliary guidance path, and training-only readouts are removed. The measured forward-only cost is reported below under the same input shape and batch protocol for both checkpoints.

Table 4: Forward-only inference on an RTX 4090 with BF16 autocast and input shape $8 \times 3 \times 224 \times 224$, using 50 warm-up and 200 timed iterations. Data loading is excluded. Per-image time is batch time divided by eight. Peak memory is allocated GPU memory in GiB.

Method	Params (M)	Batch (ms)	ms/image	Images/s	Peak (GiB)
DINOv3-L + LoRA	304.338	18.179	2.272	440.06	1.212
RELIC (4096-D)	357.872	19.941	2.493	401.18	1.430

B Proof for the Implemented Classifier

The main-text projection identities follow from the SVD and Pythagoras. For the affine CLS result, write $S = U_r \Sigma_r V_r^\top$, so $S^\dagger = V_r \Sigma_r^{-1} U_r^\top$ and $\Pi_S = U_r U_r^\top$. Then $(I - \Pi_S)S = 0$, and

$$T - SA = (I - \Pi_S)T + (\Pi_S T - SA)$$

is an orthogonal decomposition. Pythagoras gives Equation 16, and $A = S^\dagger T$ attains the minimum. The rank bound follows because every column of SA lies in the column space of S.

For the calibrated target, write $R_{\mathrm{cal}} = T - SA_0$ and insert $I = \Pi_S + (I - \Pi_S)$:

$$R_{\mathrm{cal}} = (I - \Pi_S)T + (\Pi_S T - SA_0) = R_\perp + E_{\mathrm{cal}}, \qquad E_{\mathrm{cal}} \in \mathrm{col}(S).$$

Applying $I - \Pi_S$ removes E_{cal} and gives $(I - \Pi_S)R_{\mathrm{cal}} = R_\perp$. Hence the calibrated training residual equals the orthogonal projection residual if and only if $SA_0 = \Pi_S T$.

For the patch result, $V = \Pi_S V + V_\perp$ with $S^\top V_\perp = 0$, so $\mathrm{col}([S, V]) = \mathrm{col}(S) \oplus \mathrm{col}(V_\perp)$ and its projector is $\Pi_S + \Pi_{V_\perp}$. Applying the same orthogonal decomposition to $[S, V]$ and subtracting the two residual norms gives Equation 18. The gain identity therefore concerns the orthogonal component $R_\perp$; it does not identify the in-span calibration term in R_{cal}.

The following bound establishes the stability statement for the classifier actually used by RELIC. Here Equation 40 and Equation 41 are stated in the appendix because their derivation is not needed for the main narrative.

$$|f(t_4) - f(\hat{t}_4)| \le \frac{\|w\|_2 \|\gamma\|_\infty}{\sqrt{\epsilon_{\mathrm{LN}}}} \|r_4 - D(M_4)\|_2, \tag{40}$$

$$|f(t_4) - f(\hat{t}_4)| \le \|w \odot \gamma\|_2 \, \|\psi(t_4) - \psi(\hat{t}_4)\|_2. \tag{41}$$

For a feature $z \in \mathbb{R}^d$, write the centering matrix as $H = I - \mathbf{1}\mathbf{1}^\top/d$ and the affine LayerNorm operation as

$$\mathrm{LN}(z) = \Gamma \frac{Hz}{\sqrt{\|Hz\|_2^2/d + \epsilon_{\mathrm{LN}}}} + \beta, \qquad \Gamma = \mathrm{diag}(\gamma). \tag{42}$$

Here γ, β are the classifier's LayerNorm parameters and $\epsilon_{\mathrm{LN}} > 0$ is its numerical stabilizer. Since $H^\top = H$, $H^2 = H$, and $\|H\|_2 \le 1$, it suffices to bound the derivative of $F(u) = u/\sqrt{\|u\|_2^2/d + \epsilon_{\mathrm{LN}}}$. With $a = \sqrt{\|u\|_2^2/d + \epsilon_{\mathrm{LN}}}$, differentiation gives

$$J_F(u) = a^{-1} I - \frac{uu^\top}{d\, a^3}. \tag{43}$$

For $u \neq 0$, every vector perpendicular to u is an eigenvector with eigenvalue a^{-1}. Along u, the eigenvalue is $a^{-1} - \|u\|_2^2/(da^3) = \epsilon_{\mathrm{LN}}/a^3$. Both lie in $[0, \epsilon_{\mathrm{LN}}^{-1/2}]$, since $a^2 \ge \epsilon_{\mathrm{LN}}$. At $u = 0$, the same bound holds because $J_F(0) = \epsilon_{\mathrm{LN}}^{-1/2} I$. The resulting spectral-norm bound is

$$\|J_{\mathrm{LN}}(z)\|_2 = \|\Gamma J_F(Hz) H\|_2 \le \|\gamma\|_\infty / \sqrt{\epsilon_{\mathrm{LN}}}. \tag{44}$$

Integrating the Jacobian along the line segment between arbitrary z and z' yields

$$\mathrm{LN}(z) - \mathrm{LN}(z') = \int_0^1 J_{\mathrm{LN}}(z' + \tau(z - z'))(z - z')\, d\tau, \tag{45}$$

$$\|\,\mathrm{LN}(z) - \mathrm{LN}(z')\|_2 \le \frac{\|\gamma\|_\infty}{\sqrt{\epsilon_{\mathrm{LN}}}} \|z - z'\|_2. \tag{46}$$

Applying Cauchy–Schwarz to $w^\top(\mathrm{LN}(z) - \mathrm{LN}(z'))$ and substituting $t_4 - \hat{t}_4 = r_4 - D(M_4)$ proves Equation 40. The constant is a global upper bound determined by the trained classifier and its stabilizer, and can be loose. The direct normalized-feature bound in Equation 41 is more informative when the paired features are available.

For completeness, suppose $f(t_4) > 0$ and the upper bound is less than $f(t_4)$. Then $f(\hat{t}_4) \ge f(t_4) - |f(t_4) - f(\hat{t}_4)| > 0$. If $f(t_4) < 0$, applying the same argument to $-f$ proves a negative restored score. This establishes agreement with the same teacher-feature decision at nonzero margin. It does not require a distributional model of image features or an assumption about which coordinates encode artifacts.

C Supplementary Experimental Results

C.1 Task-Specific Readout from Frozen Features

Table 6: Linear readouts of original frozen DINOv3 representations at their native dimensions. Semantic classification uses 80 classes; generated-image labels denote prompt classes. Detection follows the standard AIGIBench Setting II training protocol and tests on held-out ADM, BigGAN, Midjourney, and Wukong sources. The readout schedule is fixed in advance; all table entries are percentages.

Representation	Real semantic Acc.	Prompt-class Acc.	Detection Acc.	Detection AUC
Large CLS	97.97	84.48	78.33	92.83
7B CLS	97.77	80.42	91.90	98.92
Large patch mean	96.80	82.19	79.56	94.37
7B patch mean	97.97	78.44	92.11	99.06

We compare native CLS and mean-patch representations from frozen DINOv3-L and DINOv3-7B using linear readouts. The semantic task has 80 real-image classes; generated-image class labels are prompt labels. The forensic probe follows the fixed Setting II training protocol and is evaluated on held-out generator sources. The results show that readout performance depends on the task and token interface; they do not isolate a causal effect of distillation.

The generated-image examples compare semantic and forensic readouts on the same Chameleon inputs. Panel (a) is the input; panels (b)–(e) show DINOv3-L CLS, DINOv3-L patch mean, DINOv3-7B CLS, and DINOv3-7B patch mean readouts. The lower row reports forensic predictions and local-blur response maps over an 8×8 grid of measured regions. Blue and red indicate increases and decreases in the detector score, respectively; the probability bars use a 50% decision threshold. Across these examples, DINOv3-L CLS predicts "Real" with $P(\mathrm{AI})$ of 19.8%, 3.4%, 48.3%, and 28.2%, while the other displayed forensic readouts predict AI. These examples illustrate individual readout differences and are not an estimate of their frequency across Chameleon.

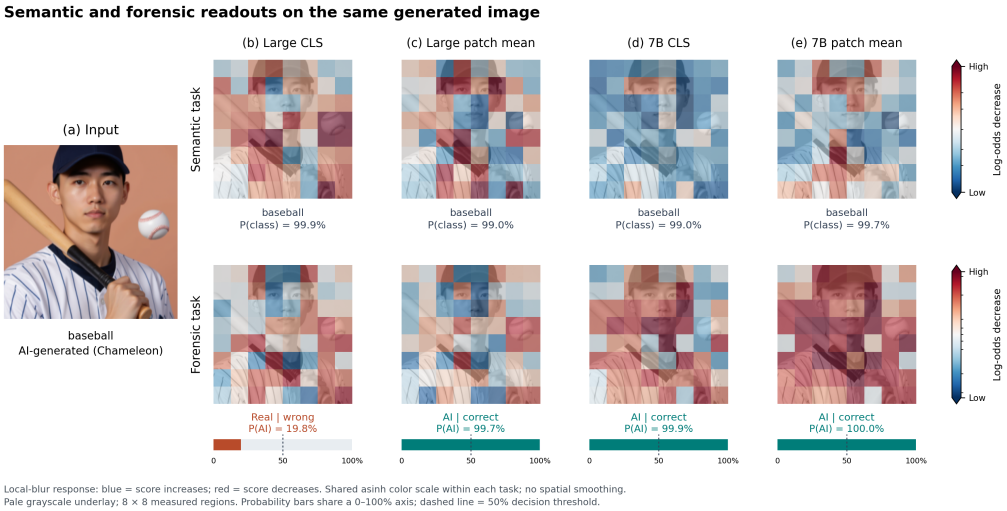

Figure 5: Task-specific semantic and forensic readouts on matched generated-image examples. Each example is evaluated with DINOv3-L CLS, DINOv3-L patch mean, DINOv3-7B CLS, and DINOv3-7B patch mean interfaces; the upper row shows semantic readouts, while the lower row shows forensic probabilities and local-blur response maps. The examples illustrate readout variation rather than dataset-level frequency.

The four readouts use the same frozen backbones and the same decision rule in each example, so the visible differences arise from the token interface rather than from a changed classifier. A semantically plausible CLS prediction can still miss a forensic cue, while a patch-based readout can respond to a localized synthesis trace. The 7B panels provide a capacity reference; they do not represent an inference-time component of RELIC.

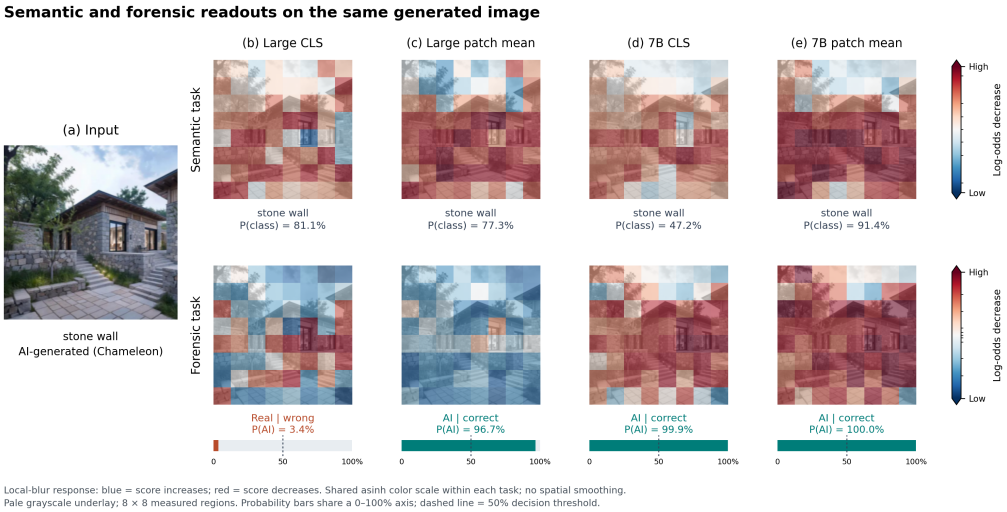

Figure 6. Stone-wall example: the Large CLS readout predicts Real (3.4% AI), while patch and 7B readouts recover the AI label with 96.7–100.0% confidence.

These panels show that semantic agreement does not imply forensic agreement: patch and 7B interfaces recover cues that Large CLS misses.

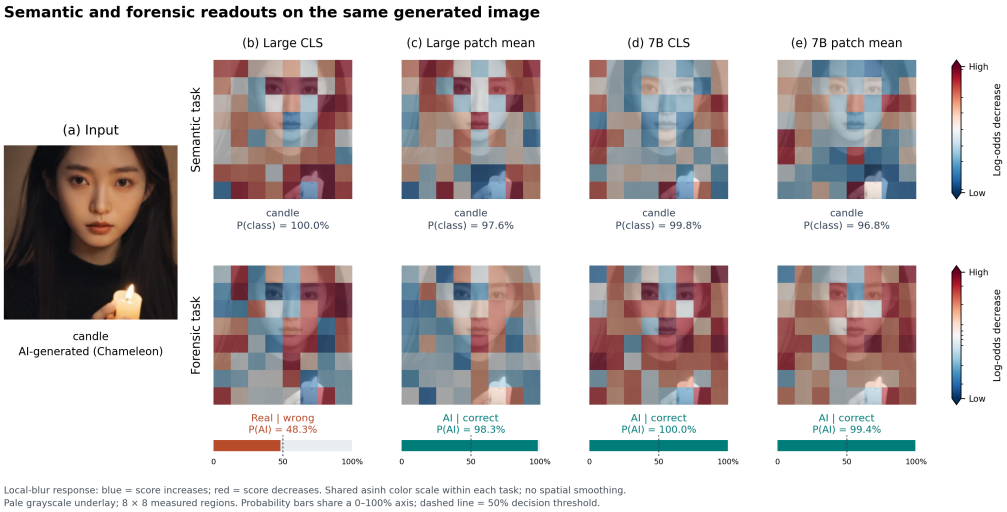

Figure 7. Candle example: the Large CLS score is ambiguous (48.3% AI), whereas patch and 7B interfaces consistently recover a high-confidence AI decision.

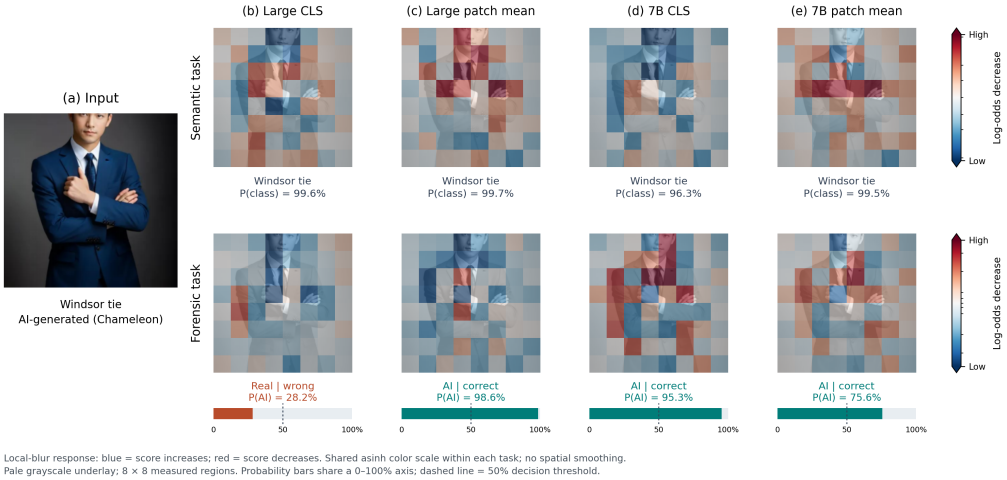

Figure 8. Windsor-tie example: the Large CLS again misses the synthetic cue (28.2% AI), while all complementary readouts cross the decision threshold, with confidence varying across interfaces.

Across the matched examples, semantic agreement and forensic agreement can therefore diverge in either direction. These panels motivate the dataset-level readouts that follow: the image examples show what a token interface can expose, whereas the tables measure how consistently that signal supports a fixed classifier over complete evaluation sets.

C.2.1 Detailed Performance

The detailed performance evidence is grouped here so that aggregate, source-wise, and operating-point measurements can be read together. All AIGIBench rows use the standard Setting II protocol and fixed checkpoints from the independent server run. The complete comparison uses the published domain order, while the generator-wise and operating-point tables use the recorded score vectors.

Table 9: Per-domain AIGIBench accuracy (%). Published baselines and RELIC follow the standard AIGIBench Setting II protocol (144K ProGAN+SD-v1.4 training images) and complete 212,802-image evaluation. The three blocks list the domains in the order used by the benchmark, and the final block also reports the domain macro-average.

Method	Venue	ProGAN	R3GAN	StyleGAN3	StyleGAN-XL	StyleSwim	WFIR	BlendFace	E4S	FaceSwap
Published baselines										
ResNet-50	CVPR'16	99.1	48.5	70.8	60.0	83.2	49.5	46.4	46.7	48.8
CNNDetection	CVPR'20	97.6	50.4	55.8	52.8	52.6	49.8	52.4	51.1	50.3
Gram-Net	CVPR'20	98.5	47.9	65.6	72.9	75.5	44.5	42.3	42.5	47.0
LGrad	CVPR'23	96.6	54.4	70.5	65.7	81.3	51.7	41.8	41.5	45.3
CLIPDetection	CVPR'23	98.4	83.5	79.6	84.6	86.4	70.0	35.0	57.0	53.1
FreqNet	AAAI'23	99.3	62.3	83.0	79.8	80.8	58.5	23.3	25.8	40.4
NPR	CVPR'24	99.4	50.8	78.4	60.3	85.7	51.6	44.5	45.0	48.1
DFFreq	arXiv'25	98.1	61.7	90.1	73.7	84.5	74.6	41.5	42.1	47.6
LaDeDa	arXiv'24	99.8	54.8	92.4	94.7	94.0	58.5	42.4	42.9	47.1
AIDE	ICLR'25	97.2	92.9	88.1	88.7	83.7	71.4	51.5	44.3	52.1
DGS-Net	ICML'26	99.7	90.5	90.7	89.6	89.8	92.3	68.1	56.9	76.4
Plug-and-play RA-Det										
RA-Det(+DINOv3-L)‡	ICML'26	100.0	94.7	72.6	97.0	91.3	83.2	51.0	52.5	51.0
RA-Det(+RELIC)‡	ICML'26	97.5	88.3	77.3	82.9	77.8	82.2	65.7	73.1	66.4
Adapted global representations										
DINOv3-L + LoRA (0.30B)	–	97.7	93.9	84.5	94.9	94.6	83.2	47.3	50.3	48.2
RELIC (0.36B)	–	96.8	91.8	92.2	93.5	92.5	83.0	58.9	70.0	61.3
DINOv3-7B + LoRA (6.72B) †	–	99.7	99.5	94.7	99.5	97.7	99.4	49.7	49.7	50.3

Method	Venue	InSwap	SimSwap	FLUX1-dev	Midjourney	GLIDE	DALLE-3	Imagen3	SD3	SDXL
Published baselines										
ResNet-50	CVPR'16	48.8	48.1	82.3	52.5	79.3	53.7	61.9	64.2	72.8
CNNDetection	CVPR'20	54.5	52.1	57.4	52.3	51.1	53.9	51.2	55.8	52.8
Gram-Net	CVPR'20	47.1	46.7	64.5	43.9	71.6	53.5	50.6	52.5	63.8
LGrad	CVPR'23	44.6	44.2	80.4	60.4	82.4	57.2	62.2	63.4	73.6
CLIPDetection	CVPR'23	43.7	43.7	80.0	65.3	76.7	75.1	78.9	84.5	84.7
FreqNet	AAAI'23	37.5	36.5	78.5	53.9	75.8	66.2	73.6	77.3	82.7
NPR	CVPR'24	47.8	47.4	95.2	68.8	82.5	57.1	85.9	91.9	86.6
DFFreq	arXiv'25	47.0	45.6	76.6	64.6	88.5	51.8	75.7	81.3	88.9
LaDeDa	arXiv'24	47.0	46.3	94.6	80.3	87.1	50.1	91.6	95.1	94.7
AIDE	ICLR'25	50.9	54.9	88.0	76.4	93.4	55.1	89.8	94.3	93.5
DGS-Net	ICML'26	78.7	83.8	88.9	73.0	92.5	41.5	89.9	91.8	92.0
Plug-and-play RA-Det										
RA-Det(+DINOv3-L)‡	ICML'26	50.6	50.3	79.7	73.6	90.6	73.8	56.1	82.2	92.0
RA-Det(+RELIC)‡	ICML'26	67.5	67.4	80.4	77.2	83.9	81.1	74.1	86.7	87.9
Adapted global representations										
DINOv3-L + LoRA (0.30B)	–	49.6	50.7	73.9	75.1	81.0	92.0	60.5	94.4	95.7
RELIC (0.36B)	–	62.7	62.7	91.9	88.8	88.1	93.8	80.1	95.2	95.5
DINOv3-7B + LoRA (6.72B) †	–	50.1	49.7	91.8	93.5	99.2	98.4	75.2	99.6	99.8

Method	Venue	BLIP	Infinite-ID	InstantID	IP-Adapter	PhotoMaker	SocialRF	CommunityAI	Mean
Published baselines									
ResNet-50	CVPR'16	99.5	49.9	61.0	62.8	49.1	55.3	52.5	61.9
CNNDetection	CVPR'20	77.2	49.7	53.2	52.0	50.1	51.1	51.3	55.1
Gram-Net	CVPR'20	98.6	50.6	75.1	54.5	50.1	52.1	52.6	58.6
LGrad	CVPR'23	93.0	50.9	72.6	70.3	59.9	53.5	55.5	62.9
CLIPDetection	CVPR'23	88.6	84.5	85.4	82.6	69.3	54.4	67.0	72.5
FreqNet	AAAI'23	93.8	79.0	79.8	78.8	77.0	54.2	55.9	66.2
NPR	CVPR'24	99.2	63.9	63.8	82.4	48.1	59.1	54.0	67.9
DFFreq	arXiv'25	97.9	70.4	91.9	83.2	88.0	57.6	54.5	71.1
LaDeDa	arXiv'24	99.0	61.5	86.5	90.8	78.4	58.6	54.5	73.4
AIDE	ICLR'25	96.4	92.2	91.8	90.0	91.7	57.8	54.1	77.6
DGS-Net	ICML'26	98.5	90.7	89.5	90.2	87.9	59.3	54.6	82.3
Plug-and-play RA-Det									
RA-Det(+DINOv3-L)‡	ICML'26	96.1	86.0	93.7	80.0	73.5	64.7	50.2	75.5
RA-Det(+RELIC)‡	ICML'26	90.8	88.5	86.1	86.2	76.2	77.1	77.2	**80.0**
Adapted global representations									
DINOv3-L + LoRA (0.30B)	–	96.6	94.4	95.8	82.3	76.5	90.3	83.1	79.5
RELIC (0.36B)	–	95.5	93.5	94.1	88.4	76.2	91.0	94.0	**85.3**
DINOv3-7B + LoRA (6.72B) †	–	99.6	99.6	99.6	96.0	84.4	96.6	95.9	86.8

† LoRA 7B CLS-only is a fine-tuned capacity reference without FTRA, distinct from the frozen teacher used to train RELIC. ‡ RA-Det reports per-domain optimal-threshold accuracy with either the original DINOv3-L or the RELIC-1024 feature interface. Published baseline numbers follow Li et al. (2025b); DGS-Net values follow its released AIGIBench re-inference log (Yan et al., 2026); RA-Det uses the local official evaluator (Wang et al., 2026). Parameter tags are approximate model sizes in B: DINOv3-L/16 is 0.30B and DINOv3-7B/16 is 6.72B (Siméoni et al., 2025); deployed RELIC is 0.36B.

The first table preserves the benchmark's three horizontal domain blocks so that published baselines, the original RA-Det interface, and the standalone RELIC representation can be compared under one

protocol. The plug-and-play RA-Det rows are kept separate from adapted global representations: the former changes only the host detector's feature interface, whereas the latter adapts a global backbone with LoRA. This distinction is important when interpreting the domain macro, because a high standalone score does not imply that a frozen host head will respond identically to a feature replacement.

The table also reports the 7B LoRA CLS-only reference as a capacity comparator. It is fine-tuned for the task and is not the teacher used by RELIC at inference. The complete rows therefore establish three comparisons at once: published forensic systems, a drop-in replacement for existing heads, and a compact global representation with a much larger capacity reference.

The domain macro is the unweighted mean over the benchmark's 25 recorded sources, so it should be read together with the complete per-domain vector. Keeping every source visible separates broad coverage from a gain concentrated in one generator and makes regressions visible. The horizontal blocks preserve the benchmark's recorded order; they are not separate training sets or selected subsets. This presentation also keeps the Setting II comparison tied to the same fixed checkpoints used by the source-wise and operating-point analyses.

For the interface comparison, the RA-Det rows retain the host detector's decision rule, whereas the adapted global rows use the standalone classifier. Reading these rows together with the full vectors distinguishes a feature replacement from a new task-adapted model and makes the capacity reference explicit before the generator-wise breakdown.

Generator-wise transfer. Table 10 reports each of the eight GenImage-test sources for both fixed checkpoints. The source split uses SD1.4: `sdv14` is the exact-version overlap and `sdv15` is a different version in the same model family. The other six source names identify different generator families or systems.

Table 10: Generator-wise GenImage-test results from fixed checkpoints. Accuracy and AUC are percentages; ΔAcc. is RELIC minus Large + LoRA in percentage points.

Source	N	Large Acc.	Large AUC	RELIC Acc.	RELIC AUC	ΔAcc.
ADM	12,000	59.77	76.06	68.38	85.50	+8.60
BigGAN	12,000	82.85	98.71	86.94	96.91	+4.09
Midjourney	12,000	68.81	95.95	85.77	96.26	+16.96
VQDM	12,000	91.03	99.28	95.62	98.98	+4.59
glide	12,000	82.81	98.81	91.07	98.10	+8.27
sdv14	12,000	99.92	100.00	99.69	99.98	-0.23
sdv15	16,000	99.88	99.99	99.60	99.95	-0.28
wukong	12,000	99.83	100.00	99.48	99.97	-0.34

Operating points and calibration. The operating-point metrics are calculated from the saved scores on each complete dataset. No target-domain threshold is optimized. TPR at fixed FPR and FPR at fixed TPR describe ranking at specified operating points; Brier score and 15-bin ECE describe probability calibration.

Table 11: ROC operating points and calibration from fixed-checkpoint predictions, pooled within each complete dataset. Rates and 15-bin ECE are percentages; Brier scores are unscaled. The RELIC AIGIBench row uses the patch-retention analysis clean forward, whose scores differ slightly from the corruption evaluation clean forward.

Dataset	Method	TPR@1%FPR	TPR@5%FPR	FPR@95%TPR	Brier	ECE
AIGIBench	Large + LoRA	29.60	57.67	34.24	0.1929	18.97
AIGIBench	RELIC	44.28	71.41	33.81	0.1328	12.49
GenImage-test	Large + LoRA	85.93	90.01	28.94	0.1317	13.73
GenImage-test	RELIC	85.70	90.84	19.67	0.0832	8.52

C.2 PLUG-AND-PLAY TRANSFER ACROSS DETECTORS

Table 7: Full 25-domain ForensicConcept accuracy (%) for the DINOv3-L + LoRA interface and the RELIC-1024 feature replacement.

Domain	DINOv3-L + LoRA	RELIC-1024	ΔAcc. (pp)
BLIP	**95.5**	**96.1**	**+0.5**
BlendFace	61.6	51.9	-9.7
CommunityAI	**69.9**	**94.8**	**+24.9**
DALLE-3	**84.0**	**95.0**	**+11.0**
E4S	68.1	55.8	-12.3
FLUX1-dev	**71.5**	**93.9**	**+22.4**
FaceSwap	58.5	52.6	-5.9
GLIDE	**76.4**	**86.7**	**+10.3**
IP-Adapter	**66.5**	**89.9**	**+23.4**
Imagen3	**53.7**	**83.9**	**+30.3**
InSwap	67.1	54.9	-12.2
Infinite-ID	**84.4**	**94.8**	**+10.3**
InstantID	95.2	95.1	-0.1
Midjourney	**69.3**	**91.5**	**+22.2**
PhotoMaker	**70.0**	**75.1**	**+5.1**
ProGAN	96.5	96.3	-0.2
R3GAN	92.0	90.5	-1.5
SD3	**88.4**	**95.8**	**+7.4**
SDXL	**93.0**	**96.1**	**+3.2**
SimSwap	66.2	55.3	-10.9
SocialRF	**81.9**	**90.9**	**+9.0**
StyleGAN-XL	93.5	92.8	-0.7
StyleGAN3	**84.1**	**91.4**	**+7.3**
StyleSwim	**90.9**	**91.4**	**+0.5**
WFIR	84.7	78.7	-6.0
25-domain macro	**78.5**	**83.7**	**+5.1**

Fixed sigmoid threshold 0.5; detector heads and concept mapping remain frozen. The macro includes Midjourney, and ΔAcc. is computed from unrounded values.

Table 8: Full 25-domain RA-Det accuracy (%) for the original DINOv3-L interface and the RELIC-1024 feature replacement.

Domain	DINOv3-L	RELIC-1024	ΔAcc. (pp)
BLIP	96.1	90.8	-5.4
BlendFace	**51.0**	**65.7**	**+14.7**
CommunityAI	**50.2**	**77.2**	**+27.0**
DALLE-3	**73.8**	**81.1**	**+7.3**
E4S	**52.5**	**73.1**	**+20.5**
FLUX1-dev	**79.7**	**80.4**	**+0.7**
FaceSwap	**51.0**	**66.4**	**+15.4**
GLIDE	90.6	83.9	-6.7
IP-Adapter	**80.0**	**86.2**	**+6.2**
Imagen3	**56.1**	**74.1**	**+18.0**
InSwap	**50.6**	**67.5**	**+16.9**
Infinite-ID	**86.0**	**88.5**	**+2.5**
InstantID	93.7	86.1	-7.6
Midjourney	**73.6**	**77.2**	**+3.6**
PhotoMaker	**73.5**	**76.2**	**+2.7**
ProGAN	100.0	97.5	-2.5
R3GAN	94.7	88.3	-6.4
SD3	**82.2**	**86.7**	**+4.5**
SDXL	92.0	87.9	-4.1
SimSwap	**50.3**	**67.4**	**+17.0**
SocialRF	**64.7**	**77.1**	**+12.4**
StyleGAN-XL	97.0	82.9	-14.0
StyleGAN3	**72.6**	**77.3**	**+4.7**
StyleSwim	91.3	77.8	-13.5
WFIR	83.2	82.2	-1.0
25-domain macro	**75.5**	**80.0**	**+4.5**

RA-Det selects an optimal threshold per domain under its official evaluator. The macro is the unweighted mean over all 25 domains; ΔAcc. uses unrounded values.

ForensicConcept rises from 78.5% to 83.7% in the 25-domain macro, with improvements on 15 domains and regressions on 10. The largest gains occur on Imagen3 (+30.3 points), CommunityAI

(+24.9), IP-Adapter (+23.4), and FLUX1-dev (+22.4), while E4S drops by 12.3 points. The macro gain therefore summarizes broader coverage but does not imply a uniform improvement for every generator.

RA-Det uses an independently optimized threshold for each domain, and its macro rises from 75.5% to 80.0%. RELIC improves 16 domains and regresses on 9; CommunityAI, E4S, and Imagen3 show the largest gains, whereas StyleGAN-XL and StyleSwim show the largest decreases. These rows isolate the effect of replacing the feature interface while preserving each host detector's decision rule.

C.3 What evidence is exposed, and how does it generalize?

We evaluate the forensic information accessible from different RELIC representations using matched linear readouts. For the intermediate frozen representations, we train independent probes with the same LayerNorm and linear-classifier capacity. We consider the Projected CLS, the Decoded residual, and a concatenation control that exposes the Projected CLS together with collector memory. For completeness, we additionally report the final Restored Token using the deployed RELIC classifier, which has the same LayerNorm-plus-linear classifier form.

On AIGIBench, the Decoded residual improves over the Projected CLS from 82.43% to 84.57%, showing that the recovered residual exposes additional discriminative evidence. The final Restored Token reaches 85.26%, the highest AIGIBench accuracy among the reported readouts. On Chameleon, the ordering differs: the Projected CLS + collector-memory control achieves the highest accuracy, indicating that the relative accessibility of the recovered evidence is dataset dependent rather than uniformly improved by every feature combination.

Table 12: Readouts of frozen RELIC representations. Intermediate representations are evaluated with independent LayerNorm(4096)+Linear(4096,2) probes of identical capacity (16,386 parameters). The Restored Token row reports the deployed RELIC classifier, which uses the same LayerNorm-plus-linear classifier form. Accuracy and AUC are percentages, pooled over each complete benchmark.

Representation	AIGI Acc.	AIGI AUC	Chameleon Acc.	Chameleon AUC
Projected CLS	82.43	89.13	91.82	96.99
Decoded residual	84.57	93.10	90.10	95.32
Projected CLS + collector memory	83.73	91.53	**92.43**	**97.17**
Restored Token (RELIC)	**85.26**	92.65	91.44	96.33

Does exact-version overlap explain the transfer pattern? We group GenImage by exact generator-version overlap and average accuracy and AUC equally across domains. The group without an exact-version overlap includes SD1.5, a different version from the SD1.4 training family, so it measures version-level rather than family-level novelty. Bootstrap intervals condition on fixed checkpoints and resample images within each domain and label.

Table 13: GenImage performance grouped by exact-version overlap with the SD1.4 training generator. The no-overlap group contains seven domains, including SD1.5 from the same model family. Accuracy and AUC are domain-macro percentages. The 95% accuracy intervals use 2,000 domain-by-label stratified image bootstrap replicates, conditional on fixed checkpoints.

Group	Method	N	Macro Acc.	95% CI	Macro AUC
Seen SD1.4	Large + LoRA	12,000	99.92	[99.88, 99.97]	100.00
Seen SD1.4	RELIC	12,000	99.69	[99.59, 99.79]	99.98
No exact-version overlap	Large + LoRA	88,000	83.57	[83.39, 83.75]	95.54
No exact-version overlap	RELIC	88,000	89.55	[89.37, 89.72]	96.53

Do collector slots aggregate different patch information? We analyze the attention patterns of the four collector slots at student blocks 5, 11, 15, and 23 to examine whether they collapse onto the same patch regions. Pairwise cosine similarity and top-10% intersection-over-union measure

the overlap between slots, while normalized entropy and effective patch count characterize how broadly each slot distributes its attention. The relatively low pairwise overlap, together with non-degenerate attention spread, indicates that the collector slots aggregate different information from the student patch tokens rather than repeatedly attending to identical locations. These statistics are purely descriptive and do not imply that highly attended patches correspond to localized or human-interpretable forensic traces.

Table 14: Attention statistics of the four collector slots over all 212,802 AIGIBench images. Block indices are zero-based. Pairwise cosine similarity and top-10% IoU measure the overlap between slots, while normalized entropy and effective patch count characterize attention spread. These statistics quantify diversity in patch aggregation and are not intended to localize semantic or forensic traces. Values are unscaled.

Block	Pair cos.	Top-10% IoU	Norm. entropy	Eff. patches	Gate mean
5	0.2386	0.0894	0.8545	99.4	0.0727
11	0.1865	0.0116	0.8593	95.4	0.0407
15	0.1810	0.0230	0.8380	86.1	0.0768
23	0.3201	0.1528	0.8450	90.7	0.1773

C.4 How robust and practical is the deployed detector?

We apply the same corruption transformations to the fixed RELIC and DINOv3-L + LoRA checkpoints over the full AIGIBench set. Accuracy and AUC are reported with the change in accuracy from each checkpoint relative to its clean result. Both checkpoints are evaluated over the complete corruption grid.

Table 15: Corruption evaluation on all 212,802 AIGIBench images using fixed checkpoints and identical transformations for both models. Accuracy and AUC are pooled percentages. ΔAcc. measures the change from the corresponding clean accuracy in percentage points.

Condition	Large Acc.	Large AUC	ΔAcc.	RELIC Acc.	RELIC AUC	ΔAcc.
Clean	79.17	91.46	0.00	85.26	93.23	0.00
JPEG Q100	78.07	89.95	-1.10	84.11	92.29	-1.15
JPEG Q90	79.18	92.90	0.01	85.44	94.10	0.18
JPEG Q80	77.14	91.75	-2.03	83.57	93.03	-1.69
JPEG Q70	73.04	88.92	-6.13	80.79	90.93	-4.47
JPEG Q60	69.29	86.45	-9.88	78.23	88.73	-7.03
Blur σ =0	79.17	91.45	0.01	85.26	93.23	0.00
Blur σ =1	79.11	91.03	-0.06	83.56	91.71	-1.70
Blur σ =2	77.65	89.87	-1.52	81.28	89.55	-3.98
Blur σ =3	76.54	90.44	-2.63	81.64	90.62	-3.62
Resize 0.5	80.14	92.62	0.98	84.86	93.61	-0.40
Resize 0.75	79.49	92.26	0.33	85.27	93.56	0.01
Resize 1	79.17	91.45	0.01	85.26	93.23	0.00
Resize 1.25	78.94	91.09	-0.23	84.96	93.06	-0.30
Resize 1.5	78.81	90.94	-0.35	84.89	93.01	-0.37
Noise $\sigma = 0.05$	71.80	87.26	-7.37	81.07	90.09	-4.19

C.5 Why direct layer-wise distillation is insufficient

Directly matching teacher features at multiple depths does not provide the stable complementary target used by RELIC. The controlled Setting II ablation below compares direct four-layer matching with joint transfer and with the staged residual-collection schedule. The results show that joint transfer can be unstable and that direct layer-wise matching remains below the staged residual objective on the complete AIGIBench evaluation.

Table 16: Controlled comparison of teacher-transfer strategies under the standard AIGIBench Setting II protocol. All five matched ablation variants use exactly the same 144K ProGAN+SD-v1.4 training set, same seed , raw DINOv3-L initialization, and evaluation on the complete 212,802-image AIGIBench set. Thus, their performance differences arise only from the transfer or reconstruction strategy. Direct key-layer matching aligns CLS features and patch–CLS residuals at four anchors. Full joint transfer jointly optimizes the four-layer, four-slot token, residual, restored-CLS, and diversity objectives from the same pretrained Large initialization, while the no-residual variant removes only the residual reconstruction objective. The staged RELIC result is reported separately as a contextual reference because it uses pre-aligned modules and a multi-stage training schedule.

Distillation setup	Acc. (%)	BAcc. (%)	Nature R. (%)	AI R. (%)
Fresh LoRA Large CLS	79.17	79.15	92.60	65.71
No teacher targets	81.33	81.32	94.97	67.67
Direct four-key-layer matching	81.10	81.09	93.67	68.51
Joint 4 layers $\times$ 4 slots	54.52	54.49	92.79	16.19
Joint transfer, no residual loss	**81.96**	**81.95**	**95.07**	**68.82**
RELIC†: staged residual collection	85.26	84.18	89.81	80.70

Bold indicates the best result among the five strictly matched ablation variants. These variants share the same training data, initialization, random seed, optimization budget, and evaluation protocol, and differ only in the transfer or reconstruction objective. Fresh LoRA Large CLS serves as the adapted CLS-only baseline; No teacher targets tests whether collector training alone is sufficient; Direct four-key-layer matching tests direct teacher–student alignment; Joint 4 layers $\times$ 4 slots optimizes all transfer and reconstruction objectives simultaneously; and Joint transfer, no residual loss removes only residual reconstruction from the joint objective. †RELIC is not a matched control and is included only as a contextual reference because it uses pre-aligned modules and a separate multi-stage schedule.

The following tables report complete plug-and-play feature-interface evaluations. Each host detector keeps its decision head and evaluation protocol while the DINOv3-L feature extractor is replaced by the RELIC-1024 interface. Bold rows mark domains where RELIC improves accuracy. ForensicConcept uses a fixed sigmoid threshold of 0.5 and includes Midjourney in its 25-domain macro; RA-Det uses an independently optimized threshold for each domain. Their macro values therefore describe separate host settings rather than a head-to-head detector comparison.

The ablation separates three effects. A1 and B1 show that adding teacher targets or directly matching four key layers gives only about 81% accuracy, so endpoint agreement alone does not expose the complementary residual. A0 jointly optimizes layer transfer, four-slot reconstruction, restored-CLS alignment, and diversity, but falls to 54.52% and recalls only 16.19% of AI images, indicating severe objective interference under the raw initialization. Removing the residual term in A5 recovers 81.96%, yet still remains below the staged RELIC result. The gap is consistent with RELIC's separation of residual-coordinate calibration, inter-layer collection, and final readout, rather than evidence that any single direct matching loss is insufficient in isolation.

C.6 Main-model readouts

The table below preserves the detailed readout evidence summarized in the main text. It separates the projected CLS, collected token, decoded residual, and final Restored Token under the stated readout protocol. The component ladder itself is shown in the main-text ablation section as a strict component ablation under the same fixed evaluation protocol.

Table 17: AIGIBench readouts of RELIC representations. A common fresh-probe protocol evaluates the intermediate frozen representations; the Restored Token row reports the deployed classifier. Accuracy and balanced accuracy are percentages; AUC/AP are on the 0–1 scale.

Representation	Acc.	Bal. Acc.	AUC	AP
Projected CLS	82.43	82.42	0.89	0.91
Collected Token	83.73	83.72	0.92	0.93
Residual correction	84.57	84.56	0.93	0.93
Restored Token	85.26	84.18	0.93	0.93

The Collected Token and residual-correction readouts expose more discriminative evidence than the projected CLS. The Restored Token classifier reaches the best AIGIBench accuracy, 85.26%. These are representation readouts, not successive training epochs or individual collector depths.

The readout values should be interpreted as measurements of accessible evidence, not as a sequence of independently trained models. Projected CLS provides the compact global baseline; adding collector memory raises accuracy to 83.73%, and the decoded residual reaches 84.57%, showing that the recovered component is useful before the final Restored Token is formed. The Restored Token reaches 85.26%, the best AIGIBench accuracy in the readout table. The component ladder separately records the strict incremental additions used by the deployed model.

These diagnostics also clarify the intended interface. The collector and decoder compress information gathered from several depths into one global token, while the plug-and-play tables above keep the host detector unchanged. Thus the representation evidence, direct-distillation control, and cross-detector results form one chain of evidence: local features contain complementary signal, staged recovery makes it usable, and the resulting token can be deployed without the teacher or LEAD.

The readouts also delimit what the representation claim does not imply. A higher score for one intermediate token is not evidence that every token is needed at inference, and the residual probe's advantage does not mean that the projected CLS is discarded. The deployed Restored Token combines both terms and achieves the best AIGIBench accuracy in the readout table; the probe rows isolate which signals are linearly accessible under the stated capacity. This separation keeps the evidence compatible with the plug-and-play results, where only the final feature interface is exchanged.

Finally, the appendix tables should be read together rather than as independent claims. Tables 12–14 show that complementary evidence is measurable and that the collector attends to diverse patches; Table 15 tests the resulting interface under corruptions; and Table 16 explains why direct layer-wise matching is not the training route used by RELIC. The resulting picture is a staged recovery procedure with explicit operating limits, rather than an assertion that scale or any single auxiliary loss is sufficient by itself.