THE EVIDENCE WITHIN: TRACING AND RECOVERING GENERATIVE EVIDENCE FOR GENERALIZABLE AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image detection has become increasingly important as synthetic images rapidly enter real information ecosystems. Recent visual foundation models (VFMs) equipped with lightweight linear heads have emerged as strong baselines for this task. However, like many previous detectors, they still suffer from fake-side generalization failures under unseen or shifted generators, such as cross-mechanism transfer, reconstruction-style fakes, and localized editing. To investigate this failure, we propose *mechanism-aligned evidence probing* (MAEP), a role-adaptive diagnostic framework that turns mechanism-specific linear heads into evidence axes for tracing CLS and patch responses across frozen DINOv3 layers. Through MAEP, we find that images predicted as real by the final image-level decision can still retain measurable fake-side evidence in intermediate layers or patch tokens, suggesting that the final-layer CLS representation used by the base detector may overlook generation traces preserved in internal token features. Motivated by this observation, we propose MIRAGE, *Multi-dimensional Internal Rarity Aggregation for Generalizable Evidence*. MIRAGE reuses a single linear head to measure fake-side traces across multiple internal dimensions, including different layers and token types, and exploits their empirical rarity with respect to real-image calibration data to more accurately identify generated images. Without additional training or trainable parameters, MIRAGE improves balanced accuracy from 81.8% to 97.3% on AIGCDetectionBenchmark and from 91.4% to 94.3% on Chameleon, with substantial gains on localized-editing and reconstruction-style fakes. These results show that generation evidence can persist internally after final-CLS failure and be recovered through real-calibrated rarity. The code will be made publicly available.

1 INTRODUCTION

AI-generated images have evolved from creative media into real information ecosystems, where high-quality synthetic images produced by modern generative models and open-source tools can spread faster than fact-checking. For example, in 2023, a fabricated image of an explosion near the Pentagon rapidly spread on social media and was reported to have briefly affected financial markets Channel NewsAsia (2023); Alba (2023). Such incidents suggest that AI-generated image detection should not only avoid falsely flagging real images, but also reliably identify fake images in open generation scenarios.

The central obstacle in current AI-generated image detection is open-scenario generalization. Detectors often perform well on generators seen during training, but degrade substantially under different generation mechanisms or deployment conditions, including VAE-based reconstruction, diffusion, GAN, and autoregressive (AR) generation Kingma & Welling (2014); Ho et al. (2020); Rombach et al. (2022); Goodfellow et al. (2014); Ramesh et al. (2021), as well as localized editing, reconstruction pipelines, and post-processing or transmission degradations Chen et al. (2024); Cai et al. (2026); Yan et al. (2025a); Li et al. (2025a). While prior methods improve overall detection accuracy, many gains remain dominated by real-side reliability, leaving fake-side generalization under open generation scenarios insufficiently addressed. As shown in Fig. 1, representative artifact-based, frequency-domain, transformation-based, data-alignment, and VFM-based detectors Wang et al. (2020); Tan

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.

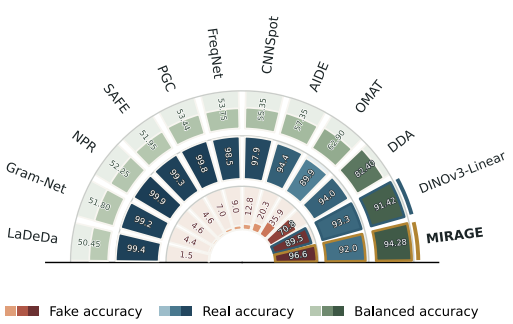


Figure 1: Fake-side collapse on Chameleon. Representative detectors often preserve high real accuracy but suffer from low fake accuracy under cross-generator shift. MIRAGE alleviates this imbalance by recovering fake samples missed by the image-level decision.

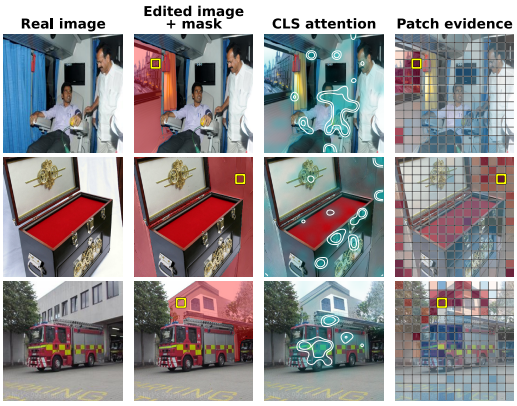


Figure 2: DINOv3-Linear failures on BR-Gen Cai et al. (2026). Columns show the real image, edited image and mask, normalized final-layer CLS-to-patch attention with high weights emphasized, and linear-head patch evidence (red: fake-like; blue: real-like). The most fake-like patch (yellow) overlaps the edited region, which receives insufficient CLS attention.

et al. (2024a); Liu et al. (2020); Tan et al. (2024b); Li et al. (2025b); Cavia et al. (2024); Chen et al. (2025b); Zhou et al. (2025; 2026a); Yan et al. (2025b); Zhou et al. (2026b) often preserve high real accuracy on Chameleon Yan et al. (2025b) but suffer from severe fake-side collapse. This real/fake imbalance also appears in strong VFM-linear baselines such as DINOv3-Linear Zhou et al. (2026b), whereas MIRAGE alleviates this collapse and improves overall generalization.

Despite the strong performance of VFM-linear detectors, their fake-side generalization can still degrade substantially under cross-mechanism shifts, localized editing, reconstruction-style pipelines, and real-world degradations. Most VFM-linear detectors, including our baseline, make predictions from the final-layer CLS representation. Recent studies have begun to uncover usable intermediate embeddings, sparse internal units, and layer-wise forensic signals in VFMs and MLLMs Huang et al. (2026); Dou et al. (2026); Lin et al. (2026). However, they leave less explored how fake-side evidence evolves across different network depths and between CLS and patch tokens, particularly when the final image-level decision fails, as well as how the surviving evidence can be exploited without substantially compromising real-side reliability.

To address this gap, we propose *mechanism-aligned evidence probing* (MAEP), a role-adaptive diagnostic framework for tracing how frozen VFMs respond to different generation mechanisms, with DINOv3 used as the representative backbone. Specifically, MAEP first constructs mechanism-controlled real/fake datasets, where the only difference between the two sides is whether a specific generation mechanism is applied. It then trains a family of candidate linear heads and uses localized forgery data to select suitable fake- and real-evidence probes. The selected probes are applied to CLS and patch tokens across all layers of frozen DINOv3, forming all-layer evidence trajectories. These trajectories reveal that generation-related evidence can emerge in intermediate layers and remain stronger or more persistent in patch tokens even when the final CLS decision predicts real.

Motivated by these observations, we propose MIRAGE, *Multi-dimensional Internal Rarity Aggregation for Generalizable Evidence*. MIRAGE reuses the base detector's linear head to measure fake-side evidence across multiple internal dimensions, including selected layers, CLS tokens, and patch tokens. It then converts such evidence into real-calibrated rarity statistics, allowing the detector to recover fake samples missed by the final image-level decision. As shown in Fig. 2, final-block CLS attention may fail to cover forged regions, whereas the highest-scoring fake-side patch selected by the linear head falls inside the edited mask in all shown cases. This alignment indicates that patch evidence can expose localized generation traces missed by the final image-level representation.

We evaluate our method on multiple AI-generated image detection benchmarks. Experiments show that MIRAGE achieves strong performance across diverse scenarios and substantially improves the fake-side generalization of a strong image-level detector under cross-mechanism shifts. On AIGCDetectionBenchmark Zhong et al. (2024), MIRAGE improves the average balanced accuracy from 81.8% to 97.3%; on Chameleon, it improves balanced accuracy from 91.4% to 94.3%. On challenging localized-editing and reconstruction-style scenarios, including BR-Gen Cai et al. (2026), DDA-COCO Chen et al. (2025b), and DRCT-inpaint (the inpainting-related subset of DRCT-2M Chen et al. (2024)), MIRAGE improves the average balanced accuracy by 11.3 points over the DINOv3-Linear baseline. Furthermore, robustness and calibration experiments show that MIRAGE can better adapt to real-domain variation by expanding clean-real calibration sources. Ablations show that the gains come from real-calibrated internal patch and layer evidence rather than simple threshold adjustment. To the best of our knowledge, this work is among the first systematic studies to trace generation-related evidence jointly across network depth and CLS/patch token granularity under final image-level detection failures, and to translate the surviving evidence into a clean-real-calibrated recovery mechanism.

Our main contributions are summarized as follows:

- We propose *mechanism-aligned evidence probing* (MAEP), a role-adaptive diagnostic framework that selects mechanism-specific fake- and real-evidence probes and traces all-layer CLS and patch evidence trajectories inside frozen VFMs.
- MAEP reveals that final image-level failure does not necessarily imply evidence disappearance: generation-related evidence can remain observable in intermediate layers and patch tokens even when the final CLS/image-level decision predicts an image as real.
- We propose MIRAGE, a real-calibrated internal rarity method that uses MAEP-motivated intermediate-layer and patch-token evidence to reformulate open-ended fake shift as multi-dimensional internal rarity testing inside the candidate-real region.
- Extensive experiments show that MIRAGE is simple: by reusing a trained linear head over internal evidence dimensions, it substantially improves detector accuracy and fake-side generalization across diverse benchmarks.

2 Related Work

Generalizable AI-generated image detection. Prior detectors exploit spatial or frequency artifacts, patch cues, reconstruction errors, and diffusion-related signals (Wang et al., 2020; Tan et al., 2024b; Zhong et al., 2024; Wang et al., 2023; Chen et al., 2024). Recent VFM-based approaches improve cross-generator generalization through linear probing, feature adaptation, or data alignment (Ojha et al., 2023; Yan et al., 2024; Chen et al., 2025b; Zhou et al., 2025; 2026b).

Internal representations for AIGI detection. Recent work explores intermediate embeddings, sparse internal units, and layer-wise forensic signals in VFMs or MLLMs (Huang et al., 2026; Dou et al., 2026; Lin et al., 2026). However, how fake-side evidence evolves jointly across depth and CLS/patch tokens when the final decision fails, and how the surviving evidence can be used while controlling real-side errors, remain underexplored. MAEP traces these dynamics, while MIRAGE converts them into clean-real-calibrated one-sided recovery.

3 Mechanism-Aligned Evidence Probing

To diagnose final-CLS detection failures, we propose *mechanism-aligned evidence probing* (MAEP), a role-adaptive framework that selects mechanism-specific evidence probes and traces their CLS and patch responses across layers of frozen VFMs. MAEP consists of controlled data construction, probe training and role selection, and all-layer trajectory analysis.

Controlled data. For each generation mechanism g, we construct a controlled full-image real/fake dataset for probe training and trajectory analysis, with matched preprocessing and postprocessing, such that the only systematic difference between the real and fake subsets is whether g is applied. We additionally construct a localized forgery dataset with edited/unedited patch annotations for probe

role selection. We follow the COCO-Inpaint-style construction Yan et al. (2025a) and use patch composition when native inpainting is unavailable.

Probe definition and role selection. We train a family of candidate linear probes on features from either frozen VFMs or LoRA-adapted variants Hu et al. (2022). Each probe can be applied to CLS and patch tokens from any layer, thereby measuring fake–real evidence across internal representations.

Given a linear probe H, let $h_{\ell,t}(x)$ denote the feature of token t at layer ℓ. The probe returns logits $z^H_{\text{fake}}(h_{\ell,t}(x))$ and $z^H_{\text{real}}(h_{\ell,t}(x))$. We define the token-level fake–real margin as

$$s^H_{\ell,t}(x) = z^H_{\text{fake}}(h_{\ell,t}(x)) - z^H_{\text{real}}(h_{\ell,t}(x)). \tag{1}$$

A larger margin indicates stronger fake-side evidence, whereas a smaller margin indicates a stronger real-side response.

To assign probe roles, we compute layer-wise fake-hit and real-hit curves on edited and unedited patches, respectively, with soft-hit curves used as auxiliary evidence. We select $H_{\text{fake}}(g)$ and $H_{\text{real}}(g)$ based on the strongest and most stable responses across layers.

Role-adaptive trajectories. After role selection, we return to the controlled full-image dataset and compute two trajectories: fake-side evidence on images generated by mechanism g, and fake–real separation between the corresponding fake and real sets.

For $r \in \{\text{cls}, \text{patch}\}$, $s^H_{\ell,r}(x)$ denotes either the CLS-token margin or the per-image mean of the eight largest patch-token margins at layer ℓ, respectively. The fake-side trajectory applies $H_{\text{fake}}(g)$ to fake images and aggregates $s^{H_{\text{fake}}(g)}_{\ell,r}(x)$ across images at each layer.

For fake–real separation, let F_g and R_g denote the controlled fake and real sets for mechanism g, respectively, and let Agg denote dataset-level mean or median aggregation.

When the same probe is selected for both roles, we write $H_s(g) = H_{\text{fake}}(g) = H_{\text{real}}(g)$ and compute the shared-head gap as

$$\begin{aligned} S_{g,r}(\ell) = &\,\text{Agg}_{x\in F_g}\, s^{H_s(g)}_{\ell,r}(x) \\ &- \text{Agg}_{x\in R_g}\, s^{H_s(g)}_{\ell,r}(x). \end{aligned} \tag{2}$$

When the two role-selected probes differ, we define the role-specific gap as

$$\begin{aligned} \Delta_{q,r}(\ell) = &\,\text{Agg}_{x\in F_g}\, s^{H_q(g)}_{\ell,r}(x) \\ &- \text{Agg}_{x\in R_g}\, s^{H_q(g)}_{\ell,r}(x), \quad q \in \{\text{fake}, \text{real}\}. \end{aligned} \tag{3}$$

For $q = \text{fake}$, this measures the additional fake-side evidence in fake images. For $q = \text{real}$, the same expression is equivalent to comparing the real-side score $-s^{H_{\text{real}}(g)}_{\ell,r}$ between real and fake images.

Because different probes may have different margin scales, we standardize the two role-specific gaps before fusion:

$$S_{g,r}(\ell) = \frac{1}{2} \sum_{q\in\{\text{fake},\text{real}\}} \frac{\Delta_{q,r}(\ell)}{\sigma^{(g)}_{q,r}}. \tag{4}$$

Here, the scale $\sigma^{(g)}_{q,r}$ in Eq. (4) is the pooled score standard deviation of probe $H_q(g)$ for evidence type r, estimated over real and fake samples and all analyzed layers of the controlled full-image dataset.

Figure 3 visualizes four representative mechanisms: VAE, SD2.1, GAN, and AR, using dataset-level medians as the main robust trend estimate. The supplementary material provides complete construction, selection, scale-estimation, trajectory, and implementation details.

These observations raise two key design questions:

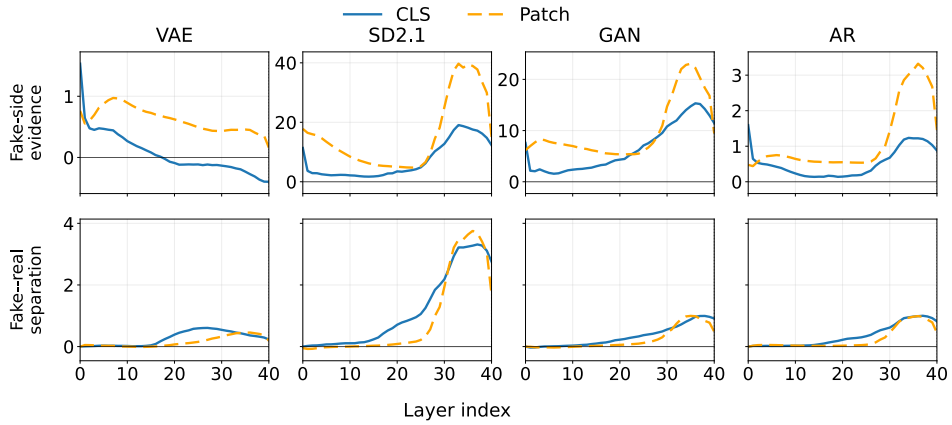


Figure 3: Layer-wise internal evidence trajectories in frozen DINOv3 obtained with mechanism-aligned evidence probing (MAEP). Curves report dataset-level medians for four representative generation mechanisms. The top row shows fake-side evidence on fake images, while the bottom row shows role-adaptive fake–real separation. Patch tokens often retain stronger evidence than CLS tokens, and the most discriminative layer is not always the final layer.

Q1: CLS and patch tokens, or CLS alone?

A: The top row of Fig. 3 shows that patch evidence is often stronger and more persistent than CLS evidence; for VAE, the final-layer CLS response even shifts toward the real side. Together with Fig. 2, this shows that patch tokens can expose localized fake traces missed by the image-level CLS representation. Our robustness analysis further shows that degradation substantially weakens fake-side evidence in the final CLS, while patch evidence remains relatively stable. Thus, the final-layer CLS alone does not provide a complete summary of generation-related evidence, motivating MIRAGE to incorporate patch tokens.

Q2: Multiple layers, or the final layer alone?

A: The bottom row of Fig. 3 shows that peak fake–real separation often occurs before the final layer, and its location varies across generation mechanisms. Therefore, no single final-layer representation uniformly captures the most discriminative evidence for all mechanisms. Intermediate layers provide complementary signals beyond the final representation, motivating MIRAGE to aggregate evidence across multiple layers.

Notably, VAE evidence emerges relatively early and weakens toward the final layer, whereas diffusion-, GAN-, and AR-related separation tends to strengthen later. This pattern is consistent with VAE reconstruction largely preserving image semantics and further illustrates that the depth at which fake evidence is most visible depends on the generation mechanism.

4 MIRAGE: Real-Calibrated Internal Rarity

Motivated by MAEP, we propose **MIRAGE**, *Multi-dimensional Internal Rarity Aggregation for Generalizable Evidence*. MIRAGE revisits candidate-real samples and tests whether their internal CLS and patch evidence exhibits calibrated rarity under a clean-real reference distribution. Figure 4 summarizes the clean-real calibration and one-sided recovery pipeline of MIRAGE.

Candidate-real region. We train a binary linear head H_0 on the final-layer CLS representation of the frozen backbone. Using the fake–real margin defined in Eq. (1), let $m_0(x) = s^{H_0}_{L,\mathrm{cls}}(x)$ denote the final-CLS margin of H_0. As shown in Fig. 4, MIRAGE keeps samples with $m_0(x) \geq 0$ as fake and applies internal rarity testing only to candidate-real samples satisfying $m_0(x) < 0$.

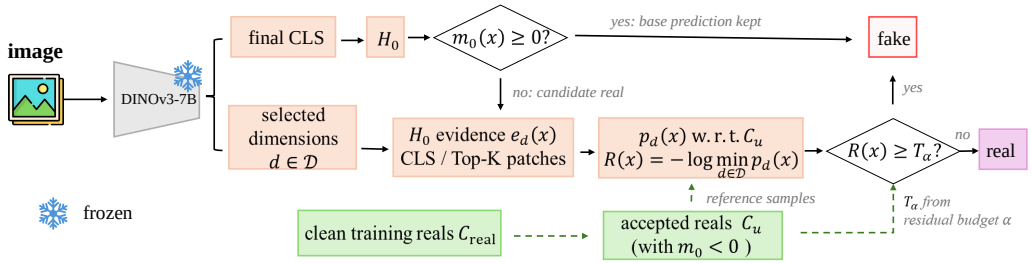


Figure 4: MIRAGE inference with clean-real calibration. The final-CLS head H_0 preserves base-fake predictions and reconsiders only candidate-real samples. Evidence from selected internal dimensions is converted into empirical upper-tail probabilities under $\mathcal{C}_{\mathcal{U}}$, aggregated into $R(x)$, and thresholded by T_α, which is calibrated from the residual clean-real budget.

Internal evidence dimensions. We reuse H_0 as an internal evidence probe over a pre-selected dimension set $\mathcal{D}$. Each $d \in \mathcal{D}$ specifies a layer and either a CLS or top-K patch statistic:

$$e_d(x) = \begin{cases} s^{H_0}_{\ell,\text{cls}}(x), & d = (\ell, \text{cls}), \\ \dfrac{1}{K}\displaystyle\sum_{j=1}^{K} s^{H_0}_{\ell,(j)}(x), & d = (\ell, \text{patch}, K). \end{cases} \tag{5}$$

Here, $s^{H_0}_{\ell,(j)}(x)$ is the j-th largest patch-token margin of the margin in Eq. (1) at layer ℓ. Thus, larger $e_d(x)$ consistently indicates stronger fake-side evidence in dimension d.

Clean-real empirical rarity. Next, we measure the rarity of these evidence scores with respect to a clean-real calibration set. To avoid data leakage, the calibration set contains only clean-real images available during training, denoted as $\mathcal{C}_{\text{real}}$, and does not use real or fake images from the test set. Because MIRAGE operates only in the candidate-real region, its reference set is $\mathcal{C}_{\mathcal{U}} = \{x' \in \mathcal{C}_{\text{real}} : m_0(x') < 0\}$, containing clean-real calibration samples accepted by the base detector. For each evidence dimension d, MIRAGE converts the evidence in Eq. (5) into a clean-real empirical upper-tail probability:

$$p_d(x) = \frac{1 + \sum_{x' \in \mathcal{C}_{\mathcal{U}}} \mathbf{1}\left[e_d(x') \geq e_d(x)\right]}{1 + |\mathcal{C}_{\mathcal{U}}|}. \tag{6}$$

A smaller $p_d(x)$ indicates that the observed fake-side evidence is more unusual under the clean-real candidate distribution.

To aggregate multiple internal evidence dimensions, MIRAGE uses an empirical min-p statistic:

$$p_{\min}(x) = \min_{d\in\mathcal{D}} p_d(x), \qquad R(x) = -\log p_{\min}(x). \tag{7}$$

The statistic $p_{\min}(x)$ retains the strongest dimension-wise rarity signal, and $R(x)$ converts it into a score for which larger values indicate greater rarity under the clean-real candidate distribution. We do not assume independence among evidence dimensions; their dependence and the multi-dimensional search effect are handled empirically by calibrating $R(x)$ on clean-real candidate samples.

Residual calibration and decision. To calibrate MIRAGE to a target clean-real rarity tolerance, we allocate only the residual rarity budget to the candidate-real region. Let α denote the target tolerance, and let α_{base} be the fraction of $\mathcal{C}_{\text{real}}$ already rejected by the base detector. Because MIRAGE acts only on $\mathcal{C}_{\mathcal{U}}$, the residual global budget is converted into the conditional gate budget:

$$\alpha_{\text{gate}} = \min\left(1, \frac{\max(\alpha - \alpha_{\text{base}}, 0)}{|\mathcal{C}_{\mathcal{U}}|/|\mathcal{C}_{\text{real}}|}\right). \tag{8}$$

We set $T_\alpha = +\infty$ when $\alpha_{\text{gate}} = 0$, which disables the rarity gate. Otherwise, the threshold is calibrated on the clean-real candidate distribution:

$$T_\alpha = \text{Quantile}_{1-\alpha_{\text{gate}}}\left(\{R(x') : x' \in \mathcal{C}_{\mathcal{U}}\}\right). \tag{9}$$

Finally, the decision rule of MIRAGE is

$$\hat{y}(x) = \begin{cases} \text{fake}, & m_0(x) \geq 0, \\ \text{fake}, & m_0(x) < 0 \text{ and } R(x) \geq T_\alpha, \\ \text{real}, & \text{otherwise.} \end{cases} \tag{10}$$

Therefore, MIRAGE keeps the base image-level classifier and its fake predictions unchanged, while adding a one-sided clean-real rarity gate within the candidate-real region. Rather than explicitly modeling or enumerating all possible fake distributions, MIRAGE evaluates whether internal evidence exhibits rarity relative to the clean-real candidate distribution. In this way, open-ended cross-mechanism fake shift is reformulated as a real-calibrated internal rarity testing problem.

5 Experiments

This section evaluates MIRAGE on standard benchmarks, challenging localized/reconstruction-style scenarios, degraded inputs, and corresponding ablations.

5.1 Experimental Setup

Datasets. We conduct a comprehensive evaluation on GenImage Zhu et al. (2023), Chameleon Yan et al. (2025b), AIGI-Now Chen et al. (2025a), AIGCDetectionBenchmark Zhong et al. (2024), BR-Gen Cai et al. (2026), DRCT-inpaint and DRCT-2M Chen et al. (2024), DDA-COCO Chen et al. (2025b), RRDataset Li et al. (2025a), and DailyBench Jiang et al. (2026). These datasets cover standard cross-generator evaluation, in-the-wild images, recent open- and closed-source generators, localized or reconstruction-style fakes, and real-world degradation.

Implementation. Main experiments freeze DINOv3-ViT-7B/16 Siméoni et al. (2025). All evaluations share a GenImage SD-v1.4-trained linear head, except AIGCDetectionBenchmark, which uses a separate ProGAN-trained head; each head is reused by MIRAGE for internal scoring. Guided by MAEP's layer-wise trajectories, MIRAGE uses patch-only $K = 1$ evidence from layers 35, 37, 38, 39, and 40, balancing generalization and computation, with the largest-margin patch selected per layer. Clean-real residual-budget calibration uses training-split real images only with $\alpha = 1\%$. All settings are fixed before testing.

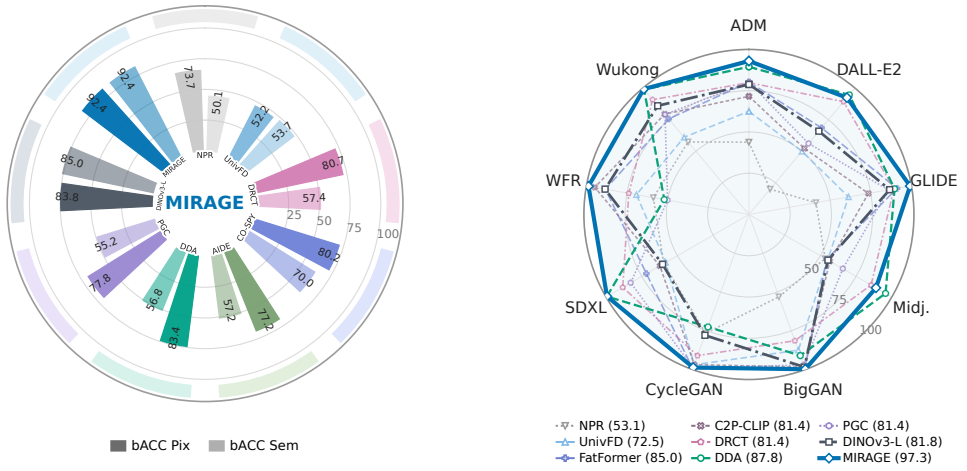
Metrics. We primarily report balanced accuracy (bACC), together with fake and real accuracy where appropriate. For the ablation study, we additionally report AUROC. For one-sided variants, AUROC uses a hierarchy-preserving ranking: base-positive samples remain above candidate-real samples, which are further ranked by internal rarity.

Compared methods. We compare against low-level artifact, frequency-domain, patch-based, reconstruction-based, and VFM-based detectors, including CNNSpot Wang et al. (2020), FreqNet Tan et al. (2024a), Gram-Net Liu et al. (2020), NPR Tan et al. (2024b), UnivFD Ojha et al. (2023), SAFE Li et al. (2025b), LaDeDa Cavia et al. (2024), DRCT Chen et al. (2024), FatFormer Liu et al. (2024), C2P-CLIP Tan et al. (2025), CO-SPY Cheng et al. (2025), B-Free Guillaro et al. (2025), Aligned Rajan et al. (2024), Effort Yan et al. (2024), DDA Chen et al. (2025b), DINO-Detect Shen et al. (2025), PGC Zhou et al. (2026a), OMAT Zhou et al. (2025), AIDE Yan et al. (2025b), RPOBE Cao et al. (2026), and FPD Jiang et al. (2026).

5.2 Results on Standard Benchmarks

Figure 5 summarizes results on two recent-generator benchmarks. On AIGI-Now, MIRAGE improves the Pix/Sem averages from $83.8/85.0\%$ to $92.4/92.4\%$. On AIGCDetectionBenchmark, it improves the macro-averaged bACC from 81.8% to 97.3%, while maintaining consistently strong performance across generators. Chameleon results are summarized in Fig. 1, where MIRAGE alleviates fake-side collapse and improves bACC from 91.4% to 94.3%.

On BR-Gen, DDA-COCO, and DRCT-inpaint, DINOv3-Linear obtains $77.8/51.9/53.1\%$ bACC, respectively. MIRAGE improves these results to $88.8/60.6/67.4\%$, while MIRAGE-Aggr. obtains



(a) AIGI-Now.

(b) AIGCDetectionBenchmark.

Figure 5: Results on standard cross-generator benchmarks. The AIGI-Now fan chart reports pixel- and semantic-level balanced accuracy, while the AIGCDetectionBenchmark radar chart reports per-generator balanced accuracy. Numbers in parentheses denote macro averages. DINOv3-L denotes our DINOv3-Linear baseline.

Table 1: DailyBench FakeBench bACC (%).

Method	SD3.5	FLUX.1	FLUX.2	Z-Image	Qwen-Img.	N.B. 2	GPT-I. 2	Avg.
NPR	62.3	70.3	63.3	67.8	70.9	61.4	76.6	67.52
Aligned	74.6	58.0	51.4	56.0	60.6	71.0	54.9	60.94
RPOBE	**92.2**	53.5	58.6	79.6	82.8	66.1	61.5	70.60
DINO-Detect	84.0	90.4	72.9	75.6	76.6	62.0	74.1	76.50
FPD	86.3	91.1	80.0	83.2	**83.3**	60.8	74.0	79.79
DINOv3-Linear	84.2	**91.2**	76.2	79.4	77.1	65.3	77.7	78.73
MIRAGE	84.9	86.6	**81.5**	**83.5**	82.3	**73.5**	**83.8**	**82.27**

87.3/65.4/74.3% at a greater real-side cost. Complete class-wise and split-level results are provided in the supplementary material.

On DailyBench Jiang et al. (2026), MIRAGE achieves 82.27% average bACC on FakeBench, outperforming FPD (79.79%) and DINOv3-Linear (78.73%). On ManipulationBench, it reaches 63.69% bACC, compared with 58.18/63.41/56.12% for DINO-Detect, FPD, and DINOv3-Linear, respectively, with positive gains across all seven manipulation generators. Additional DailyBench protocol and evaluation details are provided in the appendix.

With the backbone capacity fixed, Table 2 shows that MIRAGE achieves 98.2% bACC on GenImage, outperforming DINO-Detect (96.77%), DINOv3-Linear (96.6%), and DDA+DINOv3 (57.2%).

5.3 Robustness under Perturbations and Domain Shift

Under JPEG65 compression, MIRAGE improves fake accuracy/bACC from 78.0/88.1% to 92.8/93.6%. On RRDataset Li et al. (2025a), Table 3 shows that it improves bACC from 83.5% to 90.2% under Transfer and from 78.1% to 83.7% under Redigital degradation. These results indicate that internal patch evidence remains useful when final-CLS evidence is weakened by input degradation. Complete JPEG results and further analysis are provided in the supplementary material.

Evidence and design ablations. Table 4 compares MIRAGE with final-CLS threshold recalibration and calibrated late-layer CLS evidence. MIRAGE outperforms both variants on every benchmark and improves the mean bACC/AUROC from 81.8/90.2% to 87.8/92.5%, confirming that the

Table 2: Matched-backbone GenImage bACC (%); all methods use DINOv3-ViT-7B/16.

Method	GenImage bACC (%)
DDA + DINOv3	57.2
DINO-Detect	96.77
DINOv3-Linear	96.6
MIRAGE	**98.2**

Table 3: RRDataset bACC under real-world degradation (%).

Method	Transfer	Redigital
CNNSpot	52.9	52.7
NPR	50.0	51.9
SAFE	50.3	49.7
DDA	73.3	63.0
DINOv3-Linear	83.5	78.1
MIRAGE	**90.2**	**83.7**

Table 4: Ablations in bACC/AUROC (%). Base denotes DINOv3-Linear; Final-CLS@1% recalibrates only its decision threshold under the same 1% clean-real budget; CLS-only and MIRAGE use calibrated late-layer CLS and patch evidence, respectively. Threshold recalibration leaves AUROC unchanged.

	Base		Final-CLS@1%		CLS-only		MIRAGE	
Benchmark	bACC	AUROC	bACC	AUROC	bACC	AUROC	bACC	AUROC
Chameleon	91.4	97.7	91.0	97.7	91.8	97.5	**94.3**	**98.6**
AIGI-Now	84.4	96.5	89.8	96.5	88.5	96.6	**92.4**	**97.2**
BR-Gen	77.8	89.1	81.2	89.1	80.8	90.0	**88.8**	**94.4**
DDA-COCO	51.9	64.2	55.2	64.2	54.7	65.6	**60.6**	**69.0**
DRCT-2M	88.9	94.5	90.3	94.5	90.3	94.9	**92.4**	**95.9**
GenImage	96.6	99.5	97.4	99.5	97.3	99.5	**98.2**	**99.7**
Mean	81.8	90.2	84.2	90.2	83.9	90.7	**87.8**	**92.5**

gains arise from internal patch evidence rather than threshold adjustment alone. Final-CLS@1% leaves AUROC unchanged because it changes only the decision threshold.

Supplementary sweeps show that $\alpha = 1\%$ maximizes mean bACC; $K = 1$ exceeds $K = 20$ by 0.80 points, and the selected late-layer set exceeds L40 alone by 0.93 points. Additional GenImage, DRCT-2M, and aggregation results are provided in the supplementary material.

6 Limitations

MIRAGE targets fake-side collapse, consistent with asymmetric shortcut behavior observed in prior detectors (Ojha et al., 2023). Because its one-sided recovery can only convert candidate-real samples to fake, real-domain calibration shift may introduce false positives. Mixing real calibration images from GenImage, DDA-COCO, and BR-Gen at a $50\%/25\%/25\%$ ratio, while excluding Chameleon, improves Chameleon real accuracy/bACC from $91.99/94.28\%$ to $94.29/95.00\%$. Thus, diverse real calibration mitigates, but does not eliminate, dependence on representative real data.

7 Conclusion

MAEP reveals that generation-related evidence can persist in intermediate layers and patch tokens after the final CLS decision fails. MIRAGE converts this surviving evidence into real-calibrated rarity for one-sided fake recovery. Results across standard and shifted settings consistently attribute the gains to internal evidence rather than threshold adjustment.

AI use statement

In this work, we used generative AI tools to assist with research ideation and hypothesis refinement; feedback on experimental design; implementation, debugging, and review of data-processing, evaluation, and analysis code; generation and organization of synthetic data; formulation and checking of mathematical descriptions; literature discovery and summarization; creation and revision of figures

and tables; conversion of the manuscript to the ICLR format; and drafting, translation, and language editing. The authors made all final research decisions and manually verified all AI-assisted outputs. In particular, we checked literature and citations against the original sources, reviewed and tested AI-assisted code, recomputed reported metrics from the corresponding result files, and inspected all AI-assisted text, equations, figures, and tables for correctness. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

Ethics statement

This work studies the detection of AI-generated images, which may support content provenance and the mitigation of misinformation. Our experiments use existing research benchmarks, and we collect no new human-subject data. Because the detector remains imperfect, especially under real-domain calibration shift, its predictions should support rather than replace human verification. False positive detections may incorrectly cast doubt on authentic content and should therefore be handled cautiously in deployment.

Reproducibility statement

Sections 3 and 4 define MAEP and MIRAGE, including the evidence margins, layer-wise trajectories, empirical upper-tail probabilities, residual clean-real budget, and final decision rule. Appendix B provides the complete operational definitions, and Appendix C reports the backbone, preprocessing, training, calibration, and evaluation protocols. Appendices D and E provide complete results and additional ablations. The anonymous supplementary code archive contains the core MAEP and MIRAGE implementations, locked configurations, provenance records, and synthetic unit tests; its scope and external requirements are described in Appendix F.

References

Davey Alba. How fake AI photo of a pentagon blast went viral and briefly spooked stocks. https://www.bloomberg.com/news/articles/2023-05-22/fake-ai-photo-of-pentagon-blast-goes-viral-trips-stocks-briefly, 2023. Accessed: 2026-05-15.

Lvpan Cai, Haowei Wang, Jiayi Ji, Yanshu Zhoumen, Shen Chen, Taiping Yao, and Xiaoshuai Sun. Zooming in on fakes: A novel dataset for localized AI-generated image detection with forgery amplification approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2534–2542, 2026. doi: 10.1609/aaai.v40i4.37240.

Zijie Cao, Weijie Tu, Yao Xiao, Weijian Deng, Weiyan Chen, Liang Lin, and Pengxu Wei. Where detectors fail: Probing generative space for generalizable ai-generated image detection. In *International Conference on Machine Learning*, 2026.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024.

Channel NewsAsia. Fake image of pentagon explosion goes viral, briefly sends jitters through stock market. https://www.channelnewsasia.com/world/fake-image-pentagon-explosion-united-states-ai-generated-viral-stock-market-3507806, 2023. Accessed: 2026-05-15.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *International Conference on Machine Learning*, 2024.

Ruoxin Chen, Jiahui Gao, Kaiqing Lin, Ke-Yue Zhang, Yandan Zhao, Isabel Guan, Taiping Yao, and Shouhong Ding. Alignegemini: Generalizable ai-generated image detection through task-model alignment. *arXiv preprint arXiv:2512.06746*, 2025a. doi: 10.48550/arXiv.2512.06746.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes AI-generated image detector easier generalizable. In *Advances in Neural Information Processing Systems*, volume 38, 2025b.

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.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of AI-generated image detection with CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4356–4366, 2024.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024.

Aayush Dhakal, Subash Khanal, Srikumar Sastry, Jacob Arndt, Philipe Ambrozio Dias, Dalton Lunga, and Nathan Jacobs. SimLBR: Learning to detect fake images by learning to detect real images. *arXiv preprint arXiv:2602.20412*, 2026. doi: 10.48550/arXiv.2602.20412.

Jingtong Dou, Chuancheng Shi, Yemin Wang, Shiming Guo, Anqi Yi, Wenhua Wu, Li Zhang, Fei Shen, and Tat-Seng Chua. DNA: Uncovering universal latent forgery knowledge. *arXiv preprint arXiv:2601.22515*, 2026. doi: 10.48550/arXiv.2601.22515.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, pp. 3247–3258, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 2014.

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*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

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

Zhenhan Huang, Pin-Yu Chen, Tejaswini Pedapati, and Jianxi Gao. Intermediate representations are strong AI-generated image detectors. *arXiv preprint arXiv:2605.04358*, 2026. doi: 10.48550/arXiv.2605.04358.

Xin Jiang, Hao Tang, Junyao Gao, Zijie Yang, Meiqi Cao, Fei Shen, Dongming Zhang, Zechao Li, and Yongdong Zhang. Dailybench: A unified benchmark for ai-generated and manipulated images from modern generative models. *arXiv preprint arXiv:2607.24016*, 2026. doi: 10.48550/arXiv.2607.24016.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

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.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025b.

Kaiqing Lin, Zhiyuan Yan, Ruoxin Chen, Ke-Yue Zhang, Yue Zhou, Caiyong Piao, Bin Li, Taiping Yao, Bo Wang, Youchang Xiao, and Shouhong Ding. Deep residual injection for full-spectrum forensic signal perception in multimodal large language models. *arXiv preprint arXiv:2606.15880*, 2026. doi: 10.48550/arXiv.2606.15880.

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*, 2024.

Zhengzhe Liu, Xiaojuan Qi, and Philip H. S. Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8060–8069, 2020.

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*, pp. 24480–24489, 2023.

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

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. *arXiv preprint arXiv:2410.11835*, 2024.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Jialiang Shen, Jiyang Zheng, Yunqi Xue, Huajie Chen, Yu Yao, Hui Kang, Ruiqi Liu, Helin Gong, Yang Yang, Dadong Wang, and Tongliang Liu. DINO-Detect: A simple yet effective framework for blur-robust AI-generated image detection. *arXiv preprint arXiv:2511.12511*, 2025. doi: 10.48550/arXiv.2511.12511.

Cheng Shi, Yizhou Yu, and Sibei Yang. Vision transformers need more than registers. *arXiv preprint arXiv:2602.22394*, 2026.

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

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware Deepfake Detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024a.

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*, 2024b.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

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*, pp. 8695–8704, 2020.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Hu. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22445–22455, 2023.

Haozhen Yan, Jiahui Zhan, Yikun Ji, Yan Hong, Jun Lan, Huijia Zhu, Weiqiang Wang, and Jianfu Zhang. COCO-Inpaint: A benchmark for image inpainting detection and manipulation localization. *arXiv preprint arXiv:2504.18361*, 2025a.

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

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

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2024.

Xiaoyu Zhou, Jianwei Fei, Peipeng Yu, Jingchang Xie, Chong Cheng, and Zhihua Xia. PGC: Peak-guided calibration for generalizable AI-generated image detection. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026a.

Yue Zhou, Xinan He, Kaiqing Lin, Bing Fan, Feng Ding, and Bin Li. Breaking latent prior bias in detectors for generalizable AIGC image detection. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yue Zhou, Xinan He, Kaiqing Lin, Bing Fan, Feng Ding, and Bin Li. Simplicity prevails: The emergence of generalizable aigi detection in visual foundation models. *arXiv preprint arXiv:2602.01738*, 2026b. doi: 10.48550/arXiv.2602.01738.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Gang Lin, Wei Li, Zhijie Tu, Huajun Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In *Advances in Neural Information Processing Systems*, 2023.

A Extended Related Work

This section expands the related-work discussion in the main paper, with emphasis on the relationship between specialized forensic cues, foundation-model representations, data-centric generalization, and internal token behavior.

A.1 Specialized Forensic Representations

Early AI-generated image detectors primarily search for fingerprints introduced by particular synthesis pipelines. Spatial-domain methods exploit texture, neighboring-pixel, gradient, or noise-residual inconsistencies, whereas frequency-domain approaches identify spectral patterns caused by upsampling and other generator operations Wang et al. (2020); Frank et al. (2020); Tan et al. (2024a;b). Reconstruction-based detectors instead expose discrepancies by projecting an image through a diffusion or reconstruction model Wang et al. (2023); Chen et al. (2024).

Recognizing that weak forensic traces may be spatially sparse, later methods increasingly use patch-level or region-aware representations. SAFE introduces artifact-preserving transformations, LaDeDa performs localized real-time analysis, and BR-Gen studies challenging localized edits over object, stuff, and background regions Li et al. (2025b); Cavia et al. (2024); Cai et al. (2026). These methods establish the importance of local evidence, but their detector architectures or training objectives are explicitly designed for the forensic task. In contrast, MAEP diagnoses where generation-related evidence already exists inside a frozen VFM, and MIRAGE reuses the base detector's own linear head to recover unusually strong internal patch evidence without training an additional local forensic branch.

A.2 Foundation Models for AIGI Detection

A major line of work adapts pretrained vision foundation models to improve cross-generator generalization. UnivFD demonstrated that a lightweight classifier over frozen CLIP features transfers substantially better than detectors trained from scratch Ojha et al. (2023). Subsequent analysis showed that CLIP-based detection can remain effective with comparatively limited task-specific training data Cozzolino et al. (2024). More recent approaches, including FatFormer, Effort, DDA, and OMAT, adapt or reorganize foundation-model representations through forensic-aware transformers, subspace decomposition, aligned data, or latent-prior mitigation Liu et al. (2024); Yan et al. (2024); Chen et al. (2025b); Zhou et al. (2025).

Modern frozen VFMs equipped with a linear classifier have also emerged as strong baselines, suggesting that large-scale pretraining itself can encode useful real/fake separability Zhou et al. (2026b). However, a strong final representation does not guarantee that all useful forensic evidence is preserved by the final image-level decision. Our work therefore studies a complementary question: rather than replacing or adapting the backbone, MAEP traces evidence across layers and token types, while MIRAGE converts the surviving internal evidence into a candidate-conditioned rarity statistic.

A.3 Data-Centric Generalization and Real-Side Modeling

Training-data composition has a substantial effect on detector generalization. GenImage provides a large multi-generator benchmark, while aligned-data and bias-reduction strategies seek to prevent detectors from exploiting semantics or dataset-specific shortcuts Zhu et al. (2023); Rajan et al. (2024); Guillaro et al. (2025); Chen et al. (2025b). The AIDE/Chameleon study further demonstrates that detectors performing well on conventional benchmarks can still classify highly realistic fake images as real under strong distribution shift Yan et al. (2025b). Real-world transformations and recapturing conditions introduce additional distribution changes Li et al. (2025a).

SimLBR approaches this asymmetry by learning a compact boundary around the real-image distribution through latent blending regularization Dhakal et al. (2026). MIRAGE shares the motivation that real-side modeling is important, but differs in both objective and deployment. It does not retrain the decision boundary or synthesize latent training samples. Instead, it retains the base detector, conditions calibration on clean-real samples accepted as real by that detector, and uses the remain-

ing false-positive budget to review only candidate-real images. Thus, MIRAGE is a post-training internal-evidence calibration mechanism rather than a representation-learning regularizer.

A.4 Internal Behavior of Vision Transformers

Layer- and token-wise studies show that ViT representations are not spatially uniform. Registers were introduced to absorb high-norm artifacts in self-supervised ViTs Oquab et al. (2023); Darcet et al. (2024). More recent work argues that high-norm tokens are only one manifestation of a broader lazy aggregation behavior, in which semantically irrelevant patches can act as shortcuts for global prediction Shi et al. (2026). These studies concern representational artifacts and dense semantic alignment; they do not directly study forensic traces left by image generators.

Related layer-wise analysis has also appeared in multimodal forensic models. Deep-VRM reports that semantic information and low-level forensic signals interact differently across MLLM layers, and injects artifact-specific residual features into deeper layers Lin et al. (2026). Our setting is substantially lighter: MAEP uses linear evidence axes to analyze a frozen visual encoder, and MIRAGE requires neither an MLLM nor architectural residual injection. Together, these distinctions position our work as an internal evidence diagnosis and calibration framework rather than a new pretraining, adaptation, or feature-injection architecture.

B Additional Method Details

This section provides the full operational definitions used by mechanism-aligned evidence probing (MAEP) and MIRAGE. The descriptions below are self-contained and correspond to the final paper setting rather than earlier exploratory configurations.

B.1 Mechanism-Aligned Evidence Probing

Candidate probes. For each generation mechanism g, we train a small family of binary linear probes on controlled full-image real/fake data. Candidate probes differ in whether the frozen backbone or a LoRA-adapted backbone supplies their training features and in the amount of training data used. Once trained, every probe is applied to the post-normalization CLS and patch tokens of the frozen VFM at all analyzed layers. No backbone parameter is updated during the subsequent layer-wise diagnosis.

For a probe H, image x, layer ℓ, and token t, let $z^H_{\text{fake}}(h_{\ell,t}(x))$ and $z^H_{\text{real}}(h_{\ell,t}(x))$ be its two logits. We use the fake–real margin

$$s^H_{\ell,t}(x) = z^H_{\text{fake}}(h_{\ell,t}(x)) - z^H_{\text{real}}(h_{\ell,t}(x)). \tag{11}$$

Positive values indicate fake-side evidence, while negative values indicate real-side evidence.

Localized patch labels. Localized mask data are used only to assign probe roles; the final all-layer trajectories are computed on separate controlled full-image datasets. Each binary edit mask is resized from 224×224 to the 14×14 DINO patch grid using nearest-neighbor interpolation. For patch p, let $c_p \in [0, 1]$ be the fraction of edited pixels in its corresponding 16×16 image region. We use

$$p \in \begin{cases} \mathcal{E}(x), & c_p \geq 0.50, \\ \mathcal{U}(x), & c_p \leq 0.05, \\ \mathcal{B}(x), & 0.05 < c_p < 0.50, \end{cases} \tag{12}$$

where $\mathcal{E}(x), \mathcal{U}(x)$, and $\mathcal{B}(x)$ denote edited, unedited, and boundary patches, respectively. Boundary patches are excluded from all role-selection denominators.

Hard-hit curves. For an image with at least one valid edited or unedited patch, the per-image hard-hit rates are

$$h^H_{\text{fake}}(\ell, x) = \frac{1}{|\mathcal{E}(x)|} \sum_{p \in \mathcal{E}(x)} \mathbf{1}[s^H_{\ell,p}(x) > 0]. \tag{13}$$

$$h^H_{\text{real}}(\ell, x) = \frac{1}{|\mathcal{U}(x)|} \sum_{p \in \mathcal{U}(x)} \mathbf{1}[s^H_{\ell,p}(x) < 0]. \tag{14}$$

Table 5: Final MAEP probe-role assignment. The numbers in the probe names refer to the training-set size used by the corresponding candidate.

Mechanism	H_{fake}	H_{real}
SD2.1	Frozen-5000	LoRA-2500
VAE	LoRA-2500	LoRA-5000
GAN	Frozen-5000	LoRA-5000
AR	LoRA-2500	Frozen-2500

The dataset-level curve is an image-macro average: we first compute each valid image-level rate and then average across images. An image is omitted from a particular curve only when the corresponding patch set is empty. This prevents large edited regions from dominating the estimate merely by contributing more patches.

Soft-hit curves. Hard hits expose whether a patch falls on the expected side of the linear decision boundary. We additionally use continuous soft-hit curves as auxiliary evidence:

$$\widetilde{h}^{H}_{\text{fake}}(\ell, x) = \frac{1}{|\mathcal{E}(x)|} \sum_{p \in \mathcal{E}(x)} \sigma(s^{H}_{\ell,p}(x)). \tag{15}$$

$$\widetilde{h}^{H}_{\text{real}}(\ell, x) = \frac{1}{|\mathcal{U}(x)|} \sum_{p \in \mathcal{U}(x)} \sigma(-s^{H}_{\ell,p}(x)). \tag{16}$$

where $\sigma(u) = 1/(1 + \exp(-u))$. These curves use the same image-macro aggregation and validity rules as Eqs. (13)– (14). Hard-hit curves are the primary diagnostic; soft-hit curves reveal response strength and persistence when several probes have similar sign accuracy.

Final role-selection protocol. For every mechanism, we plot complete edited/unedited hard-hit curves and the corresponding soft-hit curves across all layers. We then select the fake- and real-evidence probes by manual visual inspection of role-specific response strength and persistence across layers. We do not reduce the curves to a single automatic score, impose a fixed layer window, or use a numerical tie-breaking rule. The final selections are listed in Table 5.

Role-adaptive all-layer trajectories. After role selection, we return to the controlled full-image real/fake sets R_g and F_g. For evidence type $r \in \{\text{cls}, \text{patch}\}$, let $s^{H}_{\ell,r}(x)$ denote the CLS margin or, for patch evidence, the per-image mean of the eight largest patch-token margins at layer ℓ. Dataset-level median trajectories are used as the primary robust summary, and mean trajectories are retained as supplementary checks.

When the same probe is selected for both roles, the separation trajectory is

$$S_{g,r}(\ell) = \operatorname*{Agg}_{x \in F_g} s^{H_s(g)}_{\ell,r}(x) - \operatorname*{Agg}_{x \in R_g} s^{H_s(g)}_{\ell,r}(x), \tag{17}$$

where $H_s(g) = H_{\text{fake}}(g) = H_{\text{real}}(g)$ and Agg is the dataset-level mean or median.

When different probes are selected, we first compute the role-specific gaps

$$\begin{aligned} \Delta^{(g)}_{q,r}(\ell) \quad = \quad & \text{Agg}_{x \in F_g}\, s^{H_q(g)}_{\ell,r}(x) \\ & - \text{Agg}_{x \in R_g}\, s^{H_q(g)}_{\ell,r}(x). \end{aligned} \tag{18}$$

for $q \in \{\text{fake}, \text{real}\}$. For the real-evidence probe, this expression is equivalent to comparing the real-side score $-s$ between real and fake images. Since independently trained probes can have different margin scales, each gap is standardized by $\sigma^{(g)}_{q,r}$, the pooled standard deviation of that probe's scores over the controlled real and fake samples and all analyzed layers. The fused trajectory is

$$S_{g,r}(\ell) = \frac{1}{2} \sum_{q \in \{\text{fake},\text{real}\}} \frac{\Delta^{(g)}_{q,r}(\ell)}{\sigma^{(g)}_{q,r}}. \tag{19}$$

This normalization is used only to make the two role-specific gaps comparable; it does not alter the probe-role selection itself.

B.2 MIRAGE: REAL-CALIBRATED INTERNAL RARITY

One-sided candidate-real review. Let H_0 be the base binary linear head applied to the final post-normalized CLS token, and let

$$m_0(x) = s^{H_0}_{L,\mathrm{cls}}(x) \tag{20}$$

be its final fake–real margin. The base detector predicts fake when $m_0(x) \geq 0$. MIRAGE never reverses such a prediction. It reviews only candidate-real samples with $m_0(x) < 0$ and can change them from real to fake when their internal evidence is sufficiently rare under clean real data.

Internal evidence dimensions. For a general dimension d, the evidence score is

$$e_d(x) = \begin{cases} s^{H_0}_{\ell,\mathrm{cls}}(x), & d = (\ell, \mathrm{cls}), \\ \dfrac{1}{K}\displaystyle\sum_{j=1}^{K} s^{H_0}_{\ell,(j)}(x), & d = (\ell, \mathrm{patch}, K), \end{cases} \tag{21}$$

where $s^{H_0}_{\ell,(j)}(x)$ is the j-th largest patch-token margin at layer ℓ. Larger values always represent stronger fake-side evidence. The final paper configuration is patch-only with $K = 1$ and

$$\begin{aligned} \mathcal{D} = \{&(35, \mathrm{patch}, 1), (37, \mathrm{patch}, 1), (38, \mathrm{patch}, 1), \\ &(39, \mathrm{patch}, 1), (40, \mathrm{patch}, 1)\}. \end{aligned} \tag{22}$$

Earlier middle-layer and $K = 3$ configurations are ablations and are not the final method.

Candidate-conditioned empirical p-values. Let $\mathcal{C}_{\mathrm{real}}$ be the clean-real calibration set available during training and

$$\mathcal{C}_{\mathcal{U}} = \{x' \in \mathcal{C}_{\mathrm{real}} : m_0(x') < 0\} \tag{23}$$

be the subset accepted as real by the base detector. No target-benchmark test image is used for calibration. For each dimension, MIRAGE computes the empirical upper-tail probability

$$p_d(x) = \frac{1 + \sum_{x' \in \mathcal{C}_{\mathcal{U}}} \mathbf{1}[e_d(x') \geq e_d(x)]}{1 + |\mathcal{C}_{\mathcal{U}}|}. \tag{24}$$

The additive one in the numerator and denominator prevents a zero empirical p-value. A small $p_d(x)$ means that the observed fake-side evidence is unusually large relative to clean-real candidate samples.

Joint empirical minimum-p calibration. We summarize the multi-dimensional search by

$$p_{\min}(x) = \min_{d \in \mathcal{D}} p_d(x), \qquad R(x) = -\log p_{\min}(x). \tag{25}$$

The evidence dimensions are neither assumed nor forced to be independent. Instead, the distribution of the joint search statistic R is itself estimated on $\mathcal{C}_{\mathcal{U}}$. Consequently, the gate controls the multi-dimensional search as a whole rather than applying the same nominal cutoff independently to every layer.

Residual false-positive budget. Let $N = |\mathcal{C}_{\mathrm{real}}|$, let N_1 be the number of calibration real images already predicted fake by the base detector, and define

$$\alpha_1 = \frac{N_1}{N}. \tag{26}$$

For a desired total clean-real false-positive tolerance α, the conditional rarity-gate budget is

$$\alpha_2 = \min\left(1, \frac{\max(\alpha - \alpha_1, 0)}{1 - \alpha_1}\right). \tag{27}$$

If $\alpha_1 \geq \alpha$, then $\alpha_2 = 0$ and the rarity gate is disabled. Otherwise, the threshold is the empirical upper quantile of R over $\mathcal{C}_{\mathcal{U}}$:

$$T_\alpha = \operatorname*{Quantile}_{1-\alpha_2}\big(\{R(x') : x' \in \mathcal{C}_{\mathcal{U}}\}\big). \tag{28}$$

Equivalently, one may threshold $p_{\min}$ at the corresponding lower-tail cutoff $\tau_{\min}$. The implementation uses an inclusive empirical comparison; tied calibration values can therefore make the realized finite- sample false-positive rate differ from the nominal target by a few samples. We report the realized rate in the ablation tables.

Final decision. The complete one-sided decision rule is

$$\widehat{y}(x) = \begin{cases} \text{fake}, & m_0(x) \geq 0, \\ \text{fake}, & m_0(x) < 0 \text{ and } R(x) \geq T_\alpha, \\ \text{real}, & \text{otherwise.} \end{cases} \tag{29}$$

Thus MIRAGE preserves every base-detector fake prediction and uses internal rarity only to rescue candidate-real samples that contain unusually strong fake-side patch evidence.

Final instantiated calibration. The final model uses a frozen DINOv3-ViT-7B/16 backbone, the locked raw-CLS linear head, GenImage SD-v1.4 training/nature real images for calibration, $\alpha = 0.01$, the five dimensions in Eq. (22), and empirical minimum-p aggregation. The calibration set contains 162,000 clean-real images; the base detector rejects 95, giving $\alpha_1 = 0.00058642$ and $\alpha_2 = 0.00941910$. The resulting lower-tail minimum-p cutoff is $\tau_{\min} = 0.00491643$. With the inclusive empirical threshold and ties, the realized total calibration false-positive rate is 1.0012%, effectively matching the prescribed 1% budget.

C Implementation and Evaluation Details

C.1 Training and Model Configuration

Backbone and inputs. The main detector uses a frozen DINOv3-ViT-7B/16 backbone with 4096-dimensional token features. Inputs are resized and center-cropped to 224×224 using the backbone's native image processor, with no additional data augmentation. For internal scoring, the selected hidden states are passed through the backbone normalization (`model.norm`) before the linear head. The token layout is one CLS token, four register tokens, and a 14×14 patch grid; register tokens are excluded from patch evidence.

MIRAGE evaluation configuration. The paper-wide evaluations use the locked GenImage-trained linear head; AIGCDetectionBenchmark follows its designated ProGAN training protocol and uses the separately trained head in Table 6. In both cases, the final backbone is frozen and the same trained head is reused on internal post-normalized tokens. The final MIRAGE configuration is patch-only and uses the maximum fake-real patch margin ($K = 1$) from layers 35, 37, 38, 39, and 40. MIRAGE reviews only samples whose final-CLS margin is negative, preserves every base fake prediction, aggregates the five empirical upper-tail probabilities by their minimum, and calibrates the resulting joint statistic with the residual clean-real false-positive budget at $\alpha = 1\%$. The default paper-wide calibration set is GenImage SD-v1.4 train/nature real; AIGCDetectionBenchmark uses ProGAN train real. RealMix is an additional calibration-source experiment and is not the default configuration.

C.2 Dataset and Evaluation Protocol

The layer set, K, α, and decision rule are locked and used unchanged across all final reported evaluations. Calibration uses training-real images only: no target-benchmark test real or test fake image is used to select a threshold. For every real/fake pair, real accuracy is the fraction of real images predicted real and fake accuracy is the fraction of fake images predicted fake. Balanced accuracy is $\mathrm{bACC} = (\mathrm{RealAcc} + \mathrm{FakeAcc})/2$. A macro result is the unweighted arithmetic mean of the corresponding generator- or split-level bACC values, rather than an image-pooled accuracy. The fixed Chameleon DINOv3-Linear baseline used for paper comparisons is 93.30% real accuracy, 89.54% fake accuracy, and 91.42% bACC.

C.3 Computing Environment

Experiments were run on Linux with Python 3.10.20, PyTorch 2.11.0 (CUDA 13.0 build), torchvision 0.26.0+cu130, Transformers 5.4.0, PEFT 0.19.1, and Accelerate 1.13.0. We used NVIDIA GeForce RTX 3090 GPUs with 24 GiB memory; each worker used one GPU, and independent extraction or training workers were parallelized when resources were available. Additional recorded dependencies include NumPy 2.2.6, Pandas 2.3.3, scikit-learn 1.7.2, SciPy 1.15.3, Pillow 12.1.1, and PyArrow 24.0.0.

D COMPLETE EXPERIMENTAL RESULTS

This section provides the complete generator-level, split-level, robustness, calibration-source, and internal-evidence results omitted from the main paper due to space constraints. Unless otherwise stated, all MIRAGE results use the frozen DINOv3-ViT-7B/16 backbone, the final Patch-only $K = 1$ configuration, late layers L35/L37/L38/L39/L40, clean-real residual-budget calibration, and target rarity tolerance $\alpha = 1\%$. All reported Real, Fake, and bACC values are percentages.

D.1 COMPLETE STANDARD-BENCHMARK RESULTS

Table 8 reports the complete generator-level AIGI-Now results under both Pix and Sem settings. The corresponding macro bACC values are 92.44% and 92.39%, respectively.

Table 9 consolidates the remaining standard-benchmark breakdowns. Panel (a) expands AIGCDetectionBenchmark to all 17 generators; its unweighted macro bACC is 97.34%, corresponding to the 97.3% main-paper result. Panel (b) reports all eight GenImage generator pairs and obtains 98.15% macro bACC. Panel (c) keeps the BR-Gen aggregate Forged view separate from the five subtype views.

D.2 LOCALIZED, RECONSTRUCTION, AND INPAINTING SCENARIOS

The BR-Gen breakdown is reported in Table 9(c); the aggregate result is never averaged together with its constituent subtypes.

Table 10 reports all six DDA-COCO generation mechanisms and the three DRCT inpainting splits. DDA-COCO remains especially challenging on the fake side, whereas the high common real accuracy shows that the degradation is primarily due to missed fake samples rather than indiscriminate real-side rejection.

Table 11 reports the official 16 DRCT-2M splits. The three inpainting splits remain part of this official 16-split macro rather than defining an additional main benchmark. The resulting DRCT-2M macro Real, Fake, and bACC values are 96.86%, 87.84%, and 92.35%, respectively.

D.3 ROBUSTNESS AND CALIBRATION-SOURCE ANALYSIS

Table 12(a) gives the complete JPEG and RRDataset results. For JPEG compression, both the matched DINOv3-Linear baseline and MIRAGE are reported. Strong compression reduces baseline fake accuracy substantially, while MIRAGE recovers much of this loss. Panel (b) additionally compares the default GenImage calibration with RealMix.

Figure 6 provides the internal evidence corresponding to the robustness results. For BR-Gen under JPEG55, the median image-level margin crosses to the real side, while the Top-K patch statistics still retain positive fake–real gaps and substantially higher fake-side rates. This behavior is consistent with global aggregation and input degradation attenuating sparse localized cues more strongly than the most suspicious patch responses. MIRAGE uses these residual patch-level rarity signals only inside the candidate-real region, thereby recovering fake samples missed by the final-CLS decision.

In Table 12(b), RealMix improves Chameleon bACC by 0.72 points, mainly through a 2.30-point increase in real accuracy while fake accuracy decreases by only 0.86 points. RealMix remains an auxiliary calibration-source experiment rather than the default configuration.

D.4 ADDITIONAL MAEP EVIDENCE

Figure 7 reports the complete mean and median probe trajectories. The upper two rows show the four controlled generation mechanisms, namely VAE, SD2.1, GAN, and AR. The lower two rows additionally report benchmark-level trajectories for DDA-COCO, DRCT-inpaint, BR-Gen, and Chameleon. Mean trajectories summarize the overall dataset response, whereas medians provide a robust view of the typical sample. Their broadly consistent trends indicate that the observed layer-wise behavior is not driven only by a small number of extreme examples.

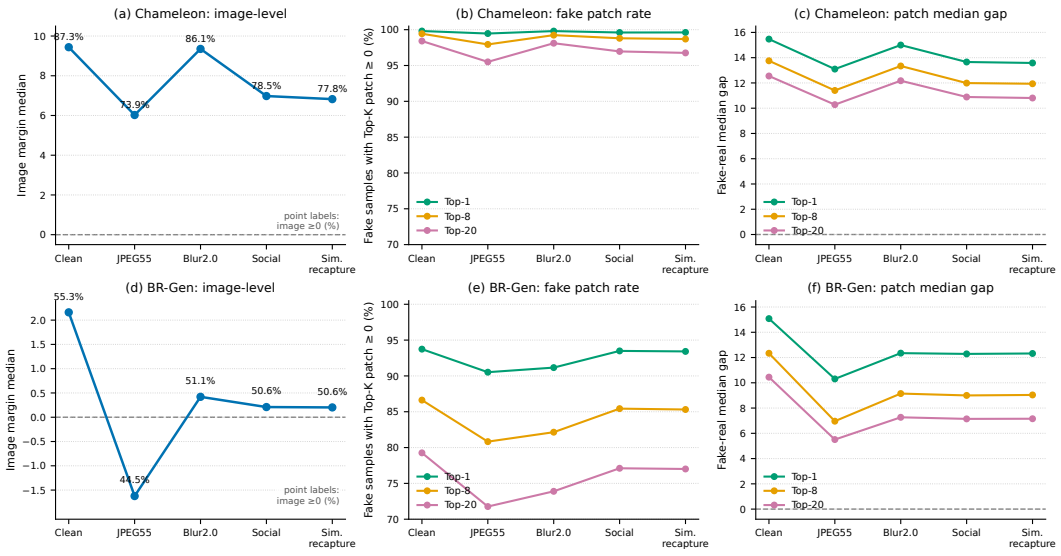


Figure 6: Image- and patch-level evidence under degradation. Image panels show the median fake margin and fake-side rate; patch panels show fake-side rate and fake–real median gaps for Top-1/8/20 statistics. Local patch evidence remains detectable when the final image-level margin weakens.

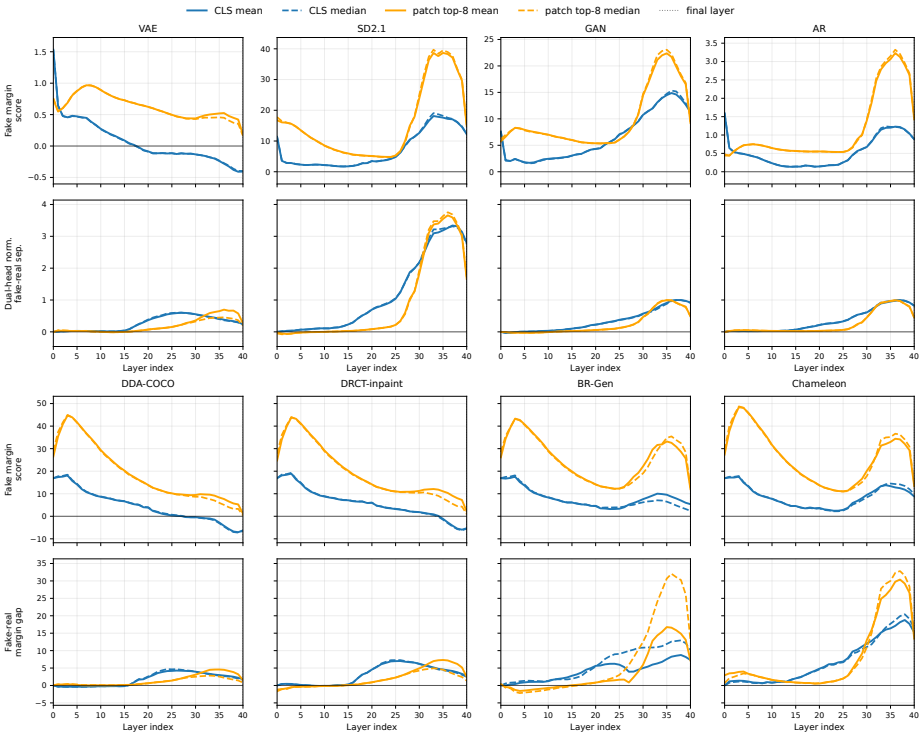
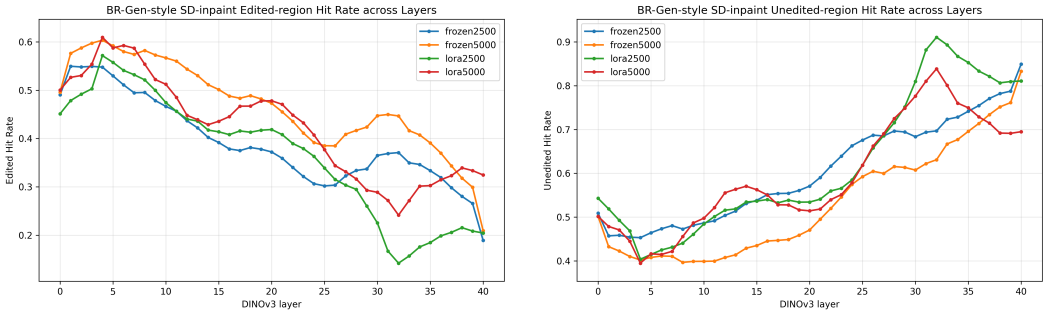


Figure 7: Complete internal-evidence trajectories. The upper two rows report the controlled VAE, SD2.1, GAN, and AR mechanisms, while the lower two rows report DDA-COCO, DRCT-inpaint, BR-Gen, and Chameleon. Solid and dashed curves denote means and medians, respectively, providing a stability check on the layer-wise evidence analysis.

Figure 8 shows a representative SD2.1 example used for assigning the fake-side and real-side probe roles. Hard-hit trajectories provide the primary visual evidence, while the corresponding soft-hit



(a) SD2.1: edited fake-side evidence

(b) SD2.1: unedited real-side evidence

Figure 8: Representative MAEP role-selection evidence for SD2.1. Edited and unedited curves expose fake-side sensitivity and real-side stability. For brevity, only the SD2.1 example is shown here; the final role assignment for all mechanisms was made by inspecting the complete hard/soft trajectories rather than by applying an automatic scalar rule.

trajectories were inspected as an auxiliary stability check. For brevity, we visualize only the SD2.1 example here; the final role assignment for all mechanisms was made from the complete curves rather than from an automatic single-number ranking.

Table 6: Final training configurations used in the experiments. A pair contains one real and one fake image. Settings not shared across a row are stated explicitly.

Component	Data / trainable parameters	Optimization	Schedule / seed	Selection
Shared MIRAGE head	GenImage SD-v1.4 train/nature versus train/AI; frozen backbone; binary $4096 \to 2$ linear head	AdamW, learning rate 10^{-3}, weight decay 0.01, batch size 128, cross-entropy	2 epochs; no learning-rate scheduler or warmup; seed 42	Highest validation balanced accuracy
AIGC head	CNNSpot/ ProGAN train real versus fake; final post-normalized CLS; binary $4096 \to 2$ linear head	AdamW, learning rate 10^{-3}, weight decay 0.01, batch size 8192	2 epochs; seed 123	Final epoch (epoch 2)
MAEP frozen probes	2500/5000 train pairs; 1000 validation and 1500 test pairs; frozen DINOv3 features and a linear probe	LR 10^{-3}; weight decay 10^{-4}; batch 512; evaluation batch 2048	At most 50 epochs; patience 8; seed 0; FP16 features	Lowest validation loss
MAEP LoRA probes	Same splits; LoRA rank 8, scale 16, dropout 0.05, plus a linear probe	LR 2×10^{-4}; weight decay 10^{-4}; batch 1; accumulation 8; evaluation batch 2	5 epochs; seed 0	Lowest validation loss

Table 7: Training, calibration, evaluation, and aggregation protocols. "Shared" denotes the locked GenImage SD-v1.4-trained head.

Dataset	Head	Calibration real	Evaluation	Aggregation
Chameleon	Shared	GenImage SD-v1.4 train real	Official real and fake test views	One real/fake pair; real accuracy, fake accuracy, and bACC
AIGI-Now	Shared	GenImage SD-v1.4 train real	Nine generators, each evaluated separately for Pix and Sem	Unweighted generator macro for Pix and Sem separately
AIGC Detection Benchmark	ProGAN-trained head	ProGAN train real	Seventeen clean-leaf generator pairs (15 unseen generators, ProGAN, and SDXL)	Per-generator bACC followed by an unweighted 17-generator macro
GenImage	Shared	GenImage SD-v1.4 train real	SD-v1.4 plus seven cross-generator validation pairs	Unweighted macro over the eight generator pairs
BR-Gen	Shared	GenImage SD-v1.4 train real	Full RealImage vs. aggregate Forged; five subtype views evaluated separately	Aggregate and subtype results remain separate
DDA-COCO	Shared	GenImage SD-v1.4 train real	COCO val2017 real paired with six fake splits	Per-split metrics; unweighted six-split bACC macro
DRCT-inpaint	Shared	GenImage SD-v1.4 train real	Three inpainting splits paired with the common val2017 real set	Auxiliary three-split macro for the hard-scenario table
DRCT-2M	Shared	GenImage SD-v1.4 train real	Official 16 fake splits, including inpainting, paired with common real	Unweighted 16-split bACC macro
RRDataset	Shared	GenImage SD-v1.4 train real	Original, Transfer, and Redigital real/fake conditions	Condition-wise reporting; no pooling
DailyBench	Shared	GenImage SD-v1.4 train real	FakeBench and ManipulationBench generator pairs	Per-generator bACC followed by an unweighted macro

Table 8: Complete AIGI-Now results for the final Patch-only $K = 1$ setting. Values are percentages. Pix and Sem are evaluated separately, and Macro is the unweighted average over all nine generators.

	Pix			Sem		
Generator	Real	Fake	bACC	Real	Fake	bACC
Nano	96.90	98.40	97.65	98.70	90.70	94.70
GPT-4o	96.20	96.40	96.30	98.90	97.70	98.30
Jimeng	96.30	83.30	89.80	98.80	92.70	95.75
Kling	96.20	95.90	96.05	98.70	88.80	93.75
Minimax	96.20	75.80	86.00	98.60	68.60	83.60
Flux Pro	96.30	87.20	91.75	99.20	94.30	96.75
Flux Krea	97.30	92.80	95.05	99.20	76.50	87.85
Flux Dev	96.90	95.10	96.00	99.20	96.10	97.65
Flux Kontext	97.20	69.50	83.35	98.70	67.60	83.15
Macro	**96.61**	**88.27**	**92.44**	**98.89**	**85.89**	**92.39**

Table 9: Complete standard-benchmark breakdowns for the final patch-only $K = 1$ setting. Panel (a) reports AIGCDetectionBenchmark bACC across 17 generators; panels (b)–(c) report Real/Fake/bACC (%) for GenImage and BR-Gen.

(a) AIGCDetectionBenchmark

Generator	bACC	Generator	bACC
ADM	93.03	SDXL	99.50
DALL-E 2	92.45	SD-v1.4	98.53
GLIDE	99.15	SD-v1.5	98.46
Midjourney	88.90	StarGAN	93.00
VQDM	99.61	StyleGAN	98.55
BigGAN	99.78	StyleGAN2	97.95
CycleGAN	98.68	WhichFaceIsReal	98.65
GauGAN	99.78	Wukong	99.21
ProGAN	99.54	**Macro**	**97.34**

(b) GenImage

Generator	Real	Fake	bACC
ADM	98.54	90.18	94.36
BigGAN	98.54	99.77	99.15
Midjourney	98.54	93.30	95.92
VQDM	98.54	99.82	99.18
GLIDE	98.54	99.33	98.94
Wukong	98.54	99.78	99.16
SD-v1.4	98.54	100.00	99.27
SD-v1.5	98.54	99.86	99.20
Macro	**98.54**	**97.76**	**98.15**

(c) BR-Gen

Type	View	Real	Fake	bACC
Aggregate	Forged	98.15	79.46	88.81
Subtype	Brush	98.15	82.53	90.34
Subtype	Power	98.15	83.67	90.91
Subtype	SDXL	98.15	69.91	84.03
Subtype	LaMa	98.15	76.01	87.08
Subtype	MAT	98.15	85.19	91.67

Table 10: Complete DDA-COCO and DRCT-inpaint split-level results for the final Patch-only $K = 1$ setting. Each fake split is paired with its common real set. DRCT-inpaint is the three-split auxiliary view used in the main hard-scenario table; the same splits remain part of the official 16-split DRCT-2M macro in Table 11.

Dataset	Fake split	Real	Fake	bACC
DDA-COCO	FLUX.1	96.86	5.89	51.38
DDA-COCO	SD-VAE EMA	96.86	27.08	61.97
DDA-COCO	SD-VAE MSE	96.86	39.72	68.29
DDA-COCO	SDXL-VAE	96.86	29.22	63.04
DDA-COCO	SD2.1	96.86	39.90	68.38
DDA-COCO	SD3.5-L	96.86	4.14	50.50
DDA macro	**6 splits**	**96.86**	**24.33**	**60.59**
DRCT-inpaint	SD2 Inpaint	96.86	43.82	70.34
DRCT-inpaint	SD Inpaint	96.86	39.76	68.31
DRCT-inpaint	SDXL Inpaint	96.86	30.34	63.60
DRCT-inpaint macro	**3 splits**	**96.86**	**37.97**	**67.42**

Table 11: Complete official DRCT-2M results for the final Patch-only $K = 1$ setting. All 16 fake splits, including the three inpainting splits, enter the unweighted macro average.

Split	Real	Fake	bACC	Split	Real	Fake	bACC
ControlNet-SDXL	96.86	99.70	98.28	SD2.1	96.86	99.66	98.26
LCM-SD1.5	96.86	99.84	98.35	SD2-Inpaint	96.86	43.82	70.34
LCM-SDXL	96.86	99.72	98.29	SD-Inpaint	96.86	39.76	68.31
LDM-256	96.86	99.96	98.41	SD1.4	96.86	99.36	98.11
ControlNet-SD2.1	96.86	99.70	98.28	SD1.5	96.86	98.96	97.91
ControlNet-SD	96.86	99.80	98.33	SDXL-Inpaint	96.86	30.34	63.60
SD-Turbo	96.86	99.84	98.35	SDXL-Base	96.86	95.80	96.33
SDXL-Turbo	96.86	99.30	98.08	SDXL-Refiner	96.86	99.92	98.39
				Macro (16 splits)	**96.86**	**87.84**	**92.35**

Table 12: Robustness and calibration-source results. Panel (a) compares the matched DINOv3-Linear baseline with the final patch-only $K = 1$ setting under JPEG compression, and reports the final setting on RRDataset transmission conditions. All results use fixed clean-real calibration at $\alpha = 1\%$. Panel (b) varies only the Chameleon calibration source.

(a) JPEG robustness and RRDataset transmission conditions

Dataset	Method	Condition	n_{real}	n_{fake}	Real	Fake	bACC
Chameleon	DINOv3-Linear	Clean	14,863	11,170	97.34	87.57	92.46
Chameleon	MIRAGE	Clean	14,863	11,170	91.99	96.56	94.28
Chameleon	DINOv3-Linear	JPEG $q = 85$	14,863	11,170	97.51	86.55	92.03
Chameleon	MIRAGE	JPEG $q = 85$	14,863	11,170	92.21	96.09	94.15
Chameleon	DINOv3-Linear	JPEG $q = 75$	14,863	11,170	97.45	87.16	92.31
Chameleon	MIRAGE	JPEG $q = 75$	14,863	11,170	92.19	96.36	94.28
Chameleon	DINOv3-Linear	JPEG $q = 65$	14,863	11,170	98.29	78.00	88.14
Chameleon	MIRAGE	JPEG $q = 65$	14,863	11,170	94.46	92.82	93.64
Chameleon	DINOv3-Linear	JPEG $q = 55$	14,863	11,170	98.50	74.49	86.50
Chameleon	MIRAGE	JPEG $q = 55$	14,863	11,170	94.92	91.14	93.03
RRDataset	MIRAGE	Original	8,500	8,500	83.06	98.27	90.67
RRDataset	MIRAGE	Transfer	8,500	8,500	94.07	86.34	90.21
RRDataset	MIRAGE	Redigital	8,499	8,500	90.15	77.20	83.68

(b) Chameleon calibration-source ablation

Detector	Calibration	Real	Fake	bACC
DINOv3-Linear	Fixed paper baseline	93.30	89.54	91.42
MIRAGE	GenImage	91.99	96.56	94.28
MIRAGE-RealMix	RealMix	94.29	95.70	95.00

E ADDITIONAL ABLATIONS

This section reports additional controlled ablations and sensitivity analyses of MIRAGE. Unless otherwise stated, all variants use the same frozen DINOv3 backbone, trained linear head, candidate-real review rule, and training-side clean-real calibration protocol.The reported default configuration is patch-only $K = 1$ evidence from L35/L37/L38/L39/L40, empirical min-p aggregation, and residual-budget calibration at $\alpha = 1\%$.

E.1 EVIDENCE SOURCES AND GATE VARIANTS

We first separate internal evidence from a direct adjustment of the image-level decision. Final-CLS@1% changes only the decision threshold of the final-layer CLS classifier. CLS-only instead applies the real-calibrated rarity gate to selected internal CLS dimensions. The default MIRAGE configuration uses only patch evidence. We also report MIRAGE-Aggr., a deliberately aggressive diagnostic variant that adds CLS evidence from L25/L26, uses top-3 patch evidence, and raises the tolerance to $\alpha = 5\%$; it is not the default method.

The final-CLS control improves fake accuracy on several benchmarks, but its real-side cost prevents it from explaining the full MIRAGE gain. CLS-only rarity further shows that internal representations contain useful evidence beyond the final image-level score. Patch-only MIRAGE provides the strongest overall real/fake balance among these controlled variants. On the three hard scenarios, MIRAGE-Aggr. recovers more fake samples but incurs a substantial additional loss of real accuracy, supporting its use only as a high-recall diagnostic configuration.

E.2 SENSITIVITY TO INTERNAL DIMENSIONS AND RARITY BUDGET

Panels (c)–(d) of Table 13 report sensitivity to the rarity tolerance, Top-K patch aggregation, and late-layer configuration. The reported MIRAGE configuration uses $\alpha = 1\%$, $K = 1$, and layers L35/L37/L38/L39/L40 consistently across the paper-wide evaluations. The sweeps below characterize sensitivity to alternative operating choices while keeping the remaining protocol unchanged.

The rarity-budget sweep illustrates the real/fake operating trade-off. At $\alpha = 1\%$, the reported configuration obtains a Mean-6 bACC of 87.766. Increasing the tolerance can further improve fake accuracy and Hard-3 bACC, but at the cost of greater real-side degradation and a lower cross-benchmark Mean-6 score.

For $K > 1$, the ablation averages the top-K patch margins within each layer; it is distinct from the historical rank-wise KMAX exploratory configuration. At the same realized calibration FPR, $K = 1$ exceeds $K = 20$ by 0.802 percentage points in Mean-6 bACC, indicating that broader patch averaging can dilute sparse fake-side evidence.

Using the five-layer L35/L37/L38/L39/L40 configuration improves Mean-6 bACC by 0.929 points and Hard-3 bACC by 2.354 points relative to L40 alone. The broader sensitivity analysis also contains nearby single-layer and leave-one-out configurations with comparable performance. These results support the usefulness and robustness of late-layer aggregation without implying that every included layer is indispensable or that this particular set is mathematically optimal.

Configuration used in the reported results. Unless otherwise stated, all final MIRAGE evaluations use the same patch-only $K = 1$ configuration over L35/L37/L38/L39/L40, empirical min-p aggregation, and residual-budget calibration at $\alpha = 1\%$.

E.3 AGGREGATION AND FALSE-POSITIVE CONTROL

Panel (e) of Table 13 compares empirical min-p with direct logical aggregation. All rarity-gate rules use the same nominal $\alpha = 1\%$, but naive OR/AND apply the cutoff independently and therefore realize different joint false-positive rates.

Naive OR obtains the highest Hard-3 bACC, but its realized calibration FPR reaches 2.103%, more than twice the nominal budget. Naive AND is conservative and uses only 0.376% calibration FPR. Empirical min-p calibrates the multi-dimensional search statistic jointly, realizes approximately

Table 13: Additional MIRAGE ablations. Benchmark cells report Real/Fake/bACC (%); Calib. FPR is the realized clean-real false-positive rate. MIRAGE-Aggr. is diagnostic rather than the default method.

(a) Evidence source and threshold controls on the six main benchmarks

Method	Calib. FPR	Cham.	AIGI-Now	BR-Gen	DDA-COCO	DRCT-2M	GenImage
Base	0.059	93.30/89.54/91.42	99.69/69.12/84.41	99.83/55.79/77.81	99.54/4.21/51.88	99.54/78.21/88.88	99.77/93.41/96.59
Final-CLS@1%	1.000	87.73/94.32/91.02	98.33/81.21/89.77	98.27/64.16/81.21	96.66/13.73/55.20	96.66/83.87/90.26	98.13/96.59/97.36
CLS-only	1.001	90.94/92.74/91.84	98.82/78.19/88.51	98.71/62.92/80.82	97.50/11.79/54.65	97.50/83.00/90.25	98.28/96.35/97.31
MIRAGE	1.001	91.99/96.56/94.28	97.74/87.08/92.41	98.15/79.46/88.81	96.86/24.33/60.59	96.86/87.84/92.35	98.54/97.76/98.15

(b) Aggressive hard-scenario diagnostic

Method	Evidence / tolerance	BR-Gen	DDA-COCO	DRCT-inpaint
Base	final CLS, threshold 0	99.83/55.79/77.81	99.54/4.21/51.88	99.54/6.64/53.09
MIRAGE	patch-only $K = 1$, $\alpha = 1\%$	98.15/79.46/88.81	96.86/24.33/60.59	96.86/37.97/67.42
MIRAGE-Aggr.	CLS L25/26 + patch $K = 3$, $\alpha = 5\%$	87.53/87.00/87.27	79.00/51.72/65.36	79.00/69.67/74.34

(c) Rarity tolerance; patch-only $K = 1$ on L35/L37/L38/L39/L40

Configuration	Calib. FPR	Mean-6			Hard-3		
		Real	Fake	bACC	Real	Fake	bACC
$\alpha = 0\%$	0.059	99.284	64.720	82.002	99.636	22.213	60.925
$\alpha = 0.1\%$	0.102	99.196	69.033	84.114	99.547	29.853	64.700
$\alpha = 0.25\%$	0.251	98.840	73.595	86.218	99.220	37.419	68.319
$\alpha = 0.5\%$	0.500	98.128	76.489	87.308	98.520	42.486	70.503
$\alpha = 1\%$	1.001	96.693	78.839	**87.766**	97.291	47.254	72.273
$\alpha = 2\%$	2.001	93.659	81.230	87.445	94.193	52.421	73.307
$\alpha = 5\%$	5.002	86.570	84.908	85.739	87.120	61.095	**74.107**
$\alpha = 10\%$	10.001	76.489	88.313	82.401	77.144	69.667	73.406
(d) Patch aggregation and late-layer accumulation at $\alpha = 1\%$							
$K = 1$	1.001	96.693	78.839	**87.766**	97.291	47.254	72.273
$K = 3$	1.001	96.497	78.737	87.617	97.024	46.490	71.757
$K = 5$	1.001	96.491	78.440	87.465	97.004	45.586	71.295
$K = 20$	1.001	96.447	77.480	86.964	97.113	42.846	69.980
L40 only	1.001	96.585	77.089	86.837	96.889	42.947	69.918
L37/L38/L39/L40	1.001	96.568	78.606	87.587	97.076	46.456	71.766
L35/L37/L38/L39/L40	1.001	96.693	78.839	**87.766**	97.291	47.254	72.273
(e) Aggregation and false-positive control at nominal $\alpha = 1\%$							
Base detector	0.058642	99.284	64.720	82.002	99.636	22.213	60.925
Naive OR	2.103086	93.339	81.417	87.378	93.878	52.858	**73.368**
Naive AND	0.375926	98.543	75.272	86.908	98.927	40.026	69.476
Empirical min-p	1.001235	96.693	78.839	**87.766**	97.291	47.254	72.273

1.001% FPR, and obtains the highest Mean-6 bACC. Its advantage is therefore budget-aware joint calibration and the overall cross-benchmark trade-off, not uniform superiority over OR on every individual benchmark.

F CODE AND DATA REPRODUCIBILITY

Release scope. The anonymous code archive provides a compact, path-independent release of the core MAEP and MIRAGE logic rather than the complete internal experiment codebase. It contains the released source modules, locked method configurations, synthetic CPU tests, and provenance records, but does not redistribute trained linear-head state dictionaries or other model weights. The core test suite is run with `python scripts/run_core_tests.py`; it checks 16 mathematical and token-handling cases without loading DINOv3, reading raw images, training a model, extracting features, or recomputing reported metrics.

Core MAEP pseudocode. Given candidate probes, post-normalization hidden states, localized masks, and controlled full-image real/fake sets:

1. Resize each localized mask to 224×224 by nearest-neighbor interpolation and partition the resulting 14×14 patch grid into edited ($c_p \geq 0.50$), unedited ($c_p \leq 0.05$), and boundary patches. Boundary patches are excluded from role-selection denominators.
2. For every candidate probe, layer, and valid patch, compute the fake–real margin. Form edited fake-evidence and unedited real-evidence hard-hit curves, together with their soft-hit counterparts, using image-macro aggregation.
3. Inspect the complete hard/soft curves across all analyzed layers and manually assign the fake-evidence and real-evidence probes for each mechanism. No automatic scalar selector, numerical tie-breaker, or fixed layer window is used.
4. Apply the selected role-specific probes to the controlled full-image sets. When the two roles use different probes, standardize their layer-wise real/fake gaps by the corresponding pooled score scales and average the two normalized gaps to obtain the role-adaptive trajectory.

The released role mapping and metric functions implement exactly this protocol; dataset construction and large-scale feature extraction are outside the distributed core.

Core MIRAGE pseudocode. Let $m_0(x)$ be the fake–real margin of the shared linear head on the final post-normalization CLS token, let $\mathcal{L} = \{35, 37, 38, 39, 40\}$, and let $\alpha = 0.01$:

1. The base detector predicts fake when $m_0(x) \geq 0$. Such a decision is final and is never reversed. The rarity gate is evaluated only when $m_0(x) < 0$.
2. Form the candidate-real calibration subset from clean training-side real images accepted by the base detector. Target-benchmark test-real images are never used for calibration.
3. At each $\ell \in \mathcal{L}$, exclude the four register tokens, begin patch tokens at index 5, and take the largest ($K = 1$) patch fake–real margin as the layer evidence.
4. Convert each layer evidence value to an empirical upper-tail p-value under the candidate-real calibration subset and aggregate the five layers by their minimum p-value.
5. Allocate only the residual clean-real false-positive budget left after the base detector and calibrate the corresponding inclusive lower-tail threshold $\tau_{\min}$.
6. For a base-detector candidate-real sample, output fake when $p_{\min}(x) \leq \tau_{\min}$ and real otherwise.

This is the locked patch-only K1 configuration. The GenImage-only calibration uses $\tau_{\min} = 0.004916432991983003$, while the weighted RealMix calibration uses $\tau_{\min} = 0.005312789579924612$ with source weights $0.50/0.25/0.25$ for GenImage, DDA-COCO, and BR-Gen.

Configuration provenance. The released archive records the locked $K = 1$ paper-wide configuration, together with its machine-readable provenance.

External assets and full reproduction. The archive does not redistribute raw datasets, trained linear-head state dictionaries, DINOv3 or generator weights, full LoRA adapters, feature banks, large score shards, the full LocalForge patch-label parquet, result-table inputs, figure inputs, caches, logs, or historical experiment code. Full end-to-end reproduction consequently requires the official DINOv3-ViT-7B/16 weights, the official datasets, substantial GPU memory and storage, and retraining or otherwise obtaining the corresponding linear heads under the disclosed training protocols. It additionally requires dataset-specific data preparation and feature-extraction orchestration. No formal 7B experiment, training run, or feature-extraction run was repeated during package assembly, and no locked paper metric was changed. The final archive is accompanied by a package-relative file manifest and SHA256 checksums for integrity verification.