

参赛学生姓名： 伍浩俊

中学： 中山市小榄中学

省份： 广东省

国家/地区： 中国

指导老师姓名： 方劲、文亮

指导老师单位： 中山市小榄中学、澳门大学

论文题目： Lightweight Cross-Generator AI

Image Detection via Complementary Forensic

Evidence

Lightweight Cross-Generator AI Image Detection via Complementary Forensic Evidence

Author: **Haojun Wu**¹

Supervisors: **Jin Fang**^{1,2} **Liang Wen**¹

¹Xiaoping Technology Innovation Lab, Zhongshan Xiaolan Senior High School,
Guangdong, 528415, China

²Faculty of Education, University of Macau, Taipa, 999078, Macau SAR, China

ABSTRACT

AI-generated images are now produced at everyday scale, which makes their detection important for multimedia forensics, but existing detectors often lose accuracy on unseen generators with different local signal statistics. We model cross-generator detection through complementary RGB and local forensic representations. Our Unified Forensic Evidence Block (UFEB) extracts signed scale, contextual, and directional responses from RGB inputs, which are fused with RGB features using gated residual injection and Hybrid2D integration; shared-private objectives keep generator-specific information away from the decision, and inference needs a single RGB image and no metadata. Under leave-one-generator-out evaluation, our 1.64M-parameter model achieves 95.63% mean AUROC and 88.47% worst-source AUROC. Ablations and detail-density analysis show that the relative contributions of RGB and UFEB vary with spatial detail organization, motivating adaptive low-false-alarm forensics. The remaining errors concentrate at low false-alarm rates on the harder sources, and the study covers six diffusion-based generators with a single training seed per configuration, so deployment is best treated as one piece of evidence in a broader forensic workflow.

Keywords: AI-generated image detection, cross-generator generalization, image forensics, local evidence, gated fusion

Contents

1	Introduction	2
2	Related Work	4
3	Method	8
3.1	Unified Forensic Evidence Block	8
3.2	RGB Block	9
3.3	Gated Residual Fusion	10
3.4	Hybrid2D Spatial Integration	11
3.5	Training Objective and Inference	12
3.6	Method Summary	14
4	Experiments	15
4.1	Experimental Setup	15
4.2	Dataset Analysis	16
4.3	Model Training	18
4.4	Cross-Generator Generalization	18
4.5	Validation-Held-out Gap	21
4.6	Ablation Studies	23
4.7	Sensitivity to the Gate Bias	26
5	Discussion	27
6	Conclusion	28
	References	30

1 Introduction

Recent advances in generative models, in both visual quality and content production capability, have brought AI-generated images (AIGIs) into widespread use across social media, digital creation, and online information dissemination, and a growing family of text-to-image systems makes both the detection of such images and the attribution of their sources a practical concern: in a systematic study of text-to-image outputs, Sha et al.¹ distinguished fake images from real ones and further attributed each fake to its source model, and found that generators share a common artifact while leaving model-specific fingerprints in their images. For information forensics, the key question is not only whether image content is plausible, but also whether the weak local traces left by the generation process can be identified from image-domain observations². These traces have a physical rather than a semantic origin, and they appear in spectral distributions, local pixel dependencies, and reconstruction behavior. Durall et al.³ showed that the up-convolution layers of generative networks fail to reproduce the spectral distributions of natural training data, and that this discrepancy alone can be used to detect generated data; Frank et al.⁴ traced the same class of frequency artifact to the upsampling operations used by all current GAN architectures and found it consistent across architectures, datasets, and resolutions. In the pixel domain, Tan et al.⁵ observed that upsampling introduces a characteristic interdependence between neighboring pixels that persists across generators, turning neighboring pixel relationships into a detector for CNN-generated images. Detector design has followed these observations in different ways, including the reconstruction error between an image and its diffusion reconstruction⁶, the reconstruction error of the autoencoder inside a latent diffusion model, which separates generated from real images without any training⁷, and a zero-shot entropy detector that needs only real images and no knowledge of generative architectures⁸.

At the same time, generative architectures continue to evolve, and training data can hardly cover all generation sources that may appear at deployment. Evaluation practice has adapted to this fact: the GenImage benchmark of Zhu et al.⁹ contains more than a million pairs of generated and real images and defines a cross-generator task, in which a detector trained on one generator is tested on the others, next to a degraded-image task that simulates low-resolution, blurred, and compressed inputs. Whether a detector can extract and transfer source-agnostic forensic structures from image signals therefore becomes central to its practical utility, and the quantity that matters is not the fit to a particular generator but the behavior of the detector under a shift of the generation process.

Existing AIGI detection methods rely mainly on two classes of signal representation. RGB-based methods learn discriminative representations from image content and overall texture. Wang et al.¹⁰ demonstrated that a standard convolutional classifier trained on a single GAN with careful pre- and post-processing, including data augmentation, generalizes surprisingly well to unseen architectures, datasets, and training methods, and argued that CNN-generated images share systematic flaws that keep them from being indistinguishable from photographs. The recipe has since moved to modern backbones, from the hierarchical Swin Transformer¹¹ to the modernized pure convolutional network ConvNeXt¹², and to large pretrained encoders such as CLIP¹³ and the self-supervised DINOv2¹⁴. Pretrained vision-language representations carry the direction further. Ojha et al.¹⁵ found that a classifier trained for real-fake discrimination is asymmetrically tuned, so that the real class becomes a sink for unseen fakes, and proposed learning-free classification in a frozen vision-language feature space through nearest neighbors or linear probing, which improves generalization to unseen diffusion and autoregressive models by a large margin. FatFormer¹⁶ argues that freezing CLIP with a linear layer leaves the representation insufficiently adapted to forgery cues, and instead adapts image features with a forgery-aware adapter that integrates local traces in the image and frequency domains, supervised by language-guided alignment; tuned on four-class ProGAN data, it reaches an average of 98% accuracy on unseen GANs and 95% on unseen diffusion models. C2P-CLIP¹⁷ decodes the detection features of CLIP into text and finds that the model separates fakes through concepts; injecting a category common prompt into the text encoder improves its detection accuracy by 12.4% over the CLIP baseline. This class of methods preserves the complete observed signal, but a rich representation also carries information that is unrelated to authenticity. Zheng et al.¹⁸ revealed that detectors suffer substantial accuracy drops when the image scene changes, attributed the failure to semantic artifacts in both real and generated images, and mitigated it by shuffling image patches before training a patch-based classifier. In face-related settings, Dong et al.¹⁹ identified implicit identity

leakage, that is, binary classifiers learn identity representations that hurt generalization, and removed it with an identity-unaware detector. Suppressing such shortcuts while keeping the content evidence remains an open problem for RGB-based detection.

The second class of methods constructs forensic or artifact representations instead of reading the image as content, following a long tradition in steganalysis, where fixed high-pass residual filters, known as rich models, expose weak local dependencies of the pixel lattice²⁰. Modern detectors reuse the principle inside learned pipelines. Luo et al.²¹ found that CNN-based face forgery detectors overfit to method-specific color textures, and fed them multi-scale high-frequency noise instead, guided by spatial attention from the residual. SPSL²² starts from the observation that upsampling is a necessary step of most face forgery techniques and that cumulative upsampling leaves obvious changes in the phase spectrum; it combines the spatial image with the phase spectrum and deliberately shallows its network so that the receptive fields stay local. The local relation network of Chen et al.²³ measures the similarity between features of local regions to form a robust similarity pattern, and fuses RGB and frequency attention for a more comprehensive representation. More recent work derives its training signal from reconstruction: LGrad²⁴ converts an image into the gradients of a pretrained network and treats them as a generalized artifact representation, with a gain of 11.4% over previous detectors; NPR⁵ builds residuals from the neighboring-pixel relationships induced by upsampling operators and evaluates on samples from 28 generative models. FerretNet²⁵ reconstructs images from neighboring pixels under a Markov random field model to expose disruptions in texture continuity and edge coherence, and detects images from more than twenty generative models with a 1.1M-parameter network trained only on four ProGAN classes. The common idea is that generation leaves a measurable deviation from natural local structure, and that measuring this deviation is a different task from recognizing content.

Two limitations are repeatedly reported for forensic representations. First, dependence on a specific forensic pattern can limit transfer. FrePGAN²⁶ attributed the overfitting of deepfake detectors to frequency-level artifacts and trained the detector together with a perturbation generator that makes generated images frequency-wise indistinguishable from real ones, so that the model moves from frequency artifacts at the beginning of training to image-level irregularities later. FreqNet²⁷ showed that frequency artifacts became generator-specific as synthesis technology advanced, and forced the detector to keep focusing on high-frequency information across spatial and channel dimensions while learning source-agnostic features from the phase and amplitude spectra, with a gain of 9.8% over previous methods on 17 GANs. Second, spatial coverage is uneven. Panoptic Patch Learning²⁸ established that every patch of a generated image carries synthetic artifacts and that using more patches improves robustness, traced the failure of naive detectors to a few-patch bias in which the model behaves like a lazy learner, and addressed it with randomized patch reconstruction and patch-wise contrastive learning. Veritas²⁹ built a dataset and evaluation protocol of unseen model architectures, emerging forgery techniques, and novel data domains, and reported that detectors which generalize across generators can still fall short on unseen forgeries and domains, which shows that a single cross-generator number does not capture deployment behavior. These observations motivate joint RGB-forensic representations. F³-Net³⁰ mines two complementary frequency-aware clues, decomposed frequency components and local frequency statistics, in a two-stream collaborative framework and achieves a clear margin on heavily compressed media. CO-SPY³¹ enhances semantic features, such as object-level cues, and pixel-artifact features, then integrates them adaptively, and reports average accuracy improvements of 11% to 34% across a benchmark of five real-image datasets and twenty-two generators. Training strategies shape what a detector relies on as well: DRCT³² treats generalizable detection as hard-sample classification, generates hard samples by high-quality diffusion reconstruction, and learns from them with contrastive training, improving cross-set accuracy by more than 10%; UCF³³ disentangles forgery-irrelevant, method-specific, and common forgery features and keeps only the common part for detection; SAFE³⁴ re-examines the training pipeline itself, removes the biases it calls weakened and overfitted artifact features by cropping instead of down-sampling, adding color and rotation augmentations, and masking patches at random to push the detector towards local evidence, and improves accuracy by 4.5% over previous methods on an open-world benchmark of 26 generative models. What remains less examined is the interaction rule itself: how much local forensic evidence should enter the content representation, at which depth

it should enter, and how the two classes of evidence should be organized so that neither dominates for the wrong reason. Fig. 1 summarizes this motivation.



Figure 1. Motivation: unseen-generator detection needs both views, fused under a learned gate.

To address the above problems, this paper proposes a dual-branch framework for cross-generator AIGI detection. The core idea is to preserve the RGB observed signal while constructing multiple classes of local forensic response, and to explicitly control the interaction between the two. Specifically, we construct a Unified Forensic Evidence Block (UFEB) that extracts online, from the raw RGB image, three classes of channel-wise signed response—scale consistency, local contextual deviation, and directional detail—so as to cover local signal variations with different spatial properties. All operators of the block are fixed and hand-defined, in the spirit of residual-based steganalysis, so UFEB itself introduces no trainable parameters and its responses cannot drift with the learned encoder. The RGB and UFEB branches remain relatively independent at shallow stages; at higher stages, gated residual fusion modulates the injection strength of the forensic responses per location and channel, and Hybrid2D then integrates local structure and cross-region dependencies. Training further adopts shared and private representation objectives³³ to constrain how authenticity-discriminative information and visible-generator information are organized; at inference, detection is performed by the shared representation alone and requires neither prompts nor generator labels.

The main contributions of this work are summarized as follows:

- We propose UFEB, which unifies scale consistency, local contextual deviation, and directional detail into a multi-response local forensic representation built entirely from fixed operators, providing a source of evidence complementary to the RGB content representation for cross-generator detection.
- We propose an RGB–UFEB dual-branch detection framework that, through gated residual fusion and Hybrid2D high-level integration, improves the stability of cross-generator detection while preserving the role of both classes of evidence; the fusion gate is learned, so the amount of forensic evidence that reaches the decision path is adapted to each image rather than fixed by design.
- We construct a systematic cross-generator evaluation protocol and validate the proposed lightweight model through extensive experiments, including branch-complementarity, structural, and response-level ablations and a detail-density analysis that connects the behavior of the model to the spatial organization of image detail.

2 Related Work

AIGI detection methods based on global visual representation learn authenticity cues directly from RGB images. This family of methods is attractive because it uses the complete observed signal and scales with standard recognition architectures: Wang et al. showed that a classifier trained on a single generator with suitable augmentation transfers

to unseen architectures and datasets, and Ojha et al. showed that a learning-free probe of a frozen vision–language feature space corrects the asymmetric tuning of a trained classifier and generalizes strongly to unseen diffusion and autoregressive models^{10,15}. The recipe also carries over to modern backbones, from the hierarchical Swin Transformer¹¹ and the modernized ConvNet¹² to CLIP¹³ and self-supervised DINOv2¹⁴ features, and task-specific adaptation adds to it: FatFormer¹⁶ adapts image features with a forgery-aware adapter over image and frequency evidence under language-guided alignment, whereas C2P-CLIP¹⁷ injects category common prompts into the text encoder. The weakness of this family is that the training objective does not tell the representation which part of the content is actually evidence: detectors can latch onto semantic artifacts that separate real and generated domains and lose accuracy when the image scene changes¹⁸, can learn identity cues that have nothing to do with forgery¹⁹, and, as the sanity check of Yan et al. reports, almost all off-the-shelf detectors misclassify genuinely challenging generated images as real, which is why their detector combines semantic embeddings with patch-level high- and low-frequency experts³⁵. This work keeps the complete observation as the output path of the network, but treats it as one of two explicitly separated evidence sources and lets a learned gate decide how much forensic evidence enters it, so that content-based shortcuts are not the only route to a decision.

Physical analysis of the generation process starts in the Fourier domain. Durall et al.³ showed that up-convolution layers fail to reproduce the spectral distribution of natural data, and Frank et al.⁴ traced the resulting frequency artifacts to the upsampling operations of GAN generators and found them consistent across architectures, datasets, and resolutions. Corvi et al.³⁶ extended the analysis to diffusion models, exposed the traces that they leave, and quantified how detectors developed for GAN images degrade on diffusion content, especially after the compression and resizing applied by social-network pipelines. Detectors built on these observations emphasize selected spectral components: SPSL²² combines the phase spectrum with a deliberately shallow spatial network, Luo et al.²¹ feed multi-scale high-frequency noise guided by residual attention, and FreqNet²⁷ keeps the network attending to high-frequency information across spatial and channel dimensions while learning source-agnostic features from the phase and amplitude spectra. The gradient view belongs to the same family: LGrad²⁴ treats the gradients of a pretrained network as a generalized artifact representation. The limitation is the stability of the selected cue: frequency artifacts became generator-specific as synthesis advanced, so detectors overfit to the artifacts of their training generators²⁷, and frequency-level perturbations can remove the cues that a detector depends on²⁶. Rather than choosing one spectral band or one filter in advance, the forensic block of this work measures several spatial orders of the signal with fixed operators, so the evidence cannot drift with training and does not depend on a band that may change with the generator.

A parallel family models local statistics instead of a global spectrum. The local relation network of Chen et al.²³ measures the similarity between features of local regions to form a robust similarity pattern, and fuses RGB and frequency attention for a more comprehensive representation. NPR⁵ builds residuals from the neighboring-pixel relationships induced by upsampling operators and evaluates them across a large set of generators, and FerretNet²⁵ reconstructs pixels under a Markov random field model to expose disruptions in texture continuity and edge coherence, detecting images from more than twenty generators with a 1.1M-parameter network. Patch-level training addresses the fact that a detector may rely on a small part of the image: Panoptic Patch Learning²⁸ shows that every patch of a generated image carries artifacts and that using more patches improves robustness, and removes the few-patch bias of naive training through randomized patch reconstruction and patch-wise contrastive learning. These methods establish that local evidence is informative and that spatial coverage matters, but each relies on one family of local measurement and inherits its failure mode. The forensic view of this work is likewise local, but measures several families at once—scale consistency, robust local context, and oriented detail—so that no single local statistic has to transfer on its own.

A third route replaces explicit measurements with the error of a reconstruction model. DIRE⁶ asks whether an image can be accurately reconstructed by a pretrained diffusion model and reports that generated images are reconstructed better than real ones, with generalization to unseen diffusion models and robustness to perturbations. AEROBLADE⁷ exploits the autoencoder that latent diffusion models use between image and latent space: generated

images are reconstructed more accurately by this component, which yields a detector that needs no training. ZED⁸ measures how surprising an image is under a model of real images learned with a lossless encoder, requiring neither generated training data nor knowledge of generative architectures. These detectors are elegant and often strong, but they depend on the availability of a suitable reconstruction model or of a probability model of real images, and their failure modes follow the component that they probe. The evidence used in this work is computed from the test image alone with fixed operators and does not assume access to a generator-specific reconstruction model.

The observation that different classes of evidence fail on different images motivates fusion. F³-Net for face forgery³⁰ learns frequency-decomposition components together with local frequency statistics in a two-stream framework and is particularly strong on heavily compressed media, while CO-SPY³¹ enhances semantic and pixel-artifact features and integrates them adaptively, reporting average accuracy gains of 11% to 34% over a benchmark of twenty-two generators. These designs demonstrate the value of complementary evidence, but they merge the two classes inside a shared transformation: the content representation has no protected identity path through the fusion, and the amount of forensic contribution is not decided per image. This work keeps the forensic stream outside the content path and injects it through a gated residual whose strength starts close to zero, so that the content representation stays intact while the injection is learned per location and channel, and the fused map is later integrated over local and row- and column-wise relations with an axial-attention block whose factorized form keeps the cost of sequence mixing manageable³⁷.

Beyond architectural fusion, the organization of the representation can itself be constrained. Domain separation networks³⁸ partition features into a shared subspace and private per-domain subspaces and show that explicitly modeling what is unique to each domain improves the shared representation; adversarial alignment with a gradient-reversal layer³⁹ pursues invariance directly; and contrastive objectives supervise representations through agreement between related samples, either between views of the same image⁴⁰ or between samples that share a label⁴¹. Deepfake detection has applied similar reasoning by decoupling common forgery features from method-specific and forgery-irrelevant factors³³. Training strategies shape the learnable signal as well: DRCT³² constructs hard samples by diffusion reconstruction and trains with contrastive learning, SAFE³⁴ rewrites the training pipeline against weakened and overfitted artifact features, and bias-free training⁴² aligns fake and real images semantically so that only generator artifacts differ. This work applies the same principle to generators: a group consistency loss aggregates the shared representations of semantically corresponding images from different generators, and shared-private objectives keep generator-specific information in the private branch, away from the representation that makes the decision.

Finally, evaluation practice decides what the reported numbers mean. GenImage⁹ contributes more than a million generated and real image pairs and defines the cross-generator task, in which a detector trained on one generator is tested on the others, next to a degraded-image task that simulates low-resolution, blurred, and compressed inputs; Corvi et al.³⁶ show that such post-processing is where detectors lose much of their accuracy. Veritas²⁹ distinguishes cross-model, cross-forgery-type, and cross-domain generalization and reports that detectors which generalize across generators can still fail on unseen forgery types and data domains, so a single average hides where the difficulty lies. The protocol used in this work follows the same spirit: six leave-one-generator-out folds, per-source results beside mean and worst-source summaries, and a detail-density analysis that links the relative contribution of the two evidence streams to the spatial organization of image detail.

Regarding the shortcomings of previous research, Table 1 summarizes the categories above together with the limitation that each family leaves for cross-generator detection. These limitations motivate the design introduced in the next section, which keeps the content representation intact and injects separately measured forensic evidence through a learned gate.

Table 1. Summary of the limitations in the cited literature. Each category corresponds to one paragraph of this section.

Category	Limitation
Global visual representations	Authenticity cues are not specified by the objective, so detectors latch onto semantic or identity cues that vanish when the scene changes; the reported behavior follows the backbone and its pretraining rather than the generation process.
Frequency and gradient artifacts	The selected spectral cue becomes generator-specific as synthesis advances, so detectors overfit to the artifacts of their training generators; frequency-level perturbations and social-network recompression can remove the cue that a detector depends on.
Local pixel dependencies	Each method measures one family of local statistics and inherits its failure mode; limited spatial coverage during training biases naive detectors toward a few dominating patches.
Reconstruction-based detection	Requires a suitable reconstruction model or a probability model of real images at inference; failure modes follow the probed component and tie the detector to known generation pipelines.
Fusion of evidence streams	The merged evidence classes share one transformation, so the content representation has no protected path through the fusion; the contribution of each class is fixed by design rather than decided per image.
Representation constraints	Constraints regularize a single representation and add no evidence that it does not already measure; invariance and disentanglement rely on generator labels and carefully weighted objectives.
Evaluation protocols	A single average over sources hides where the difficulty lies; validation accuracy is a weak predictor of held-out behavior, and dataset shortcuts are rarely audited.

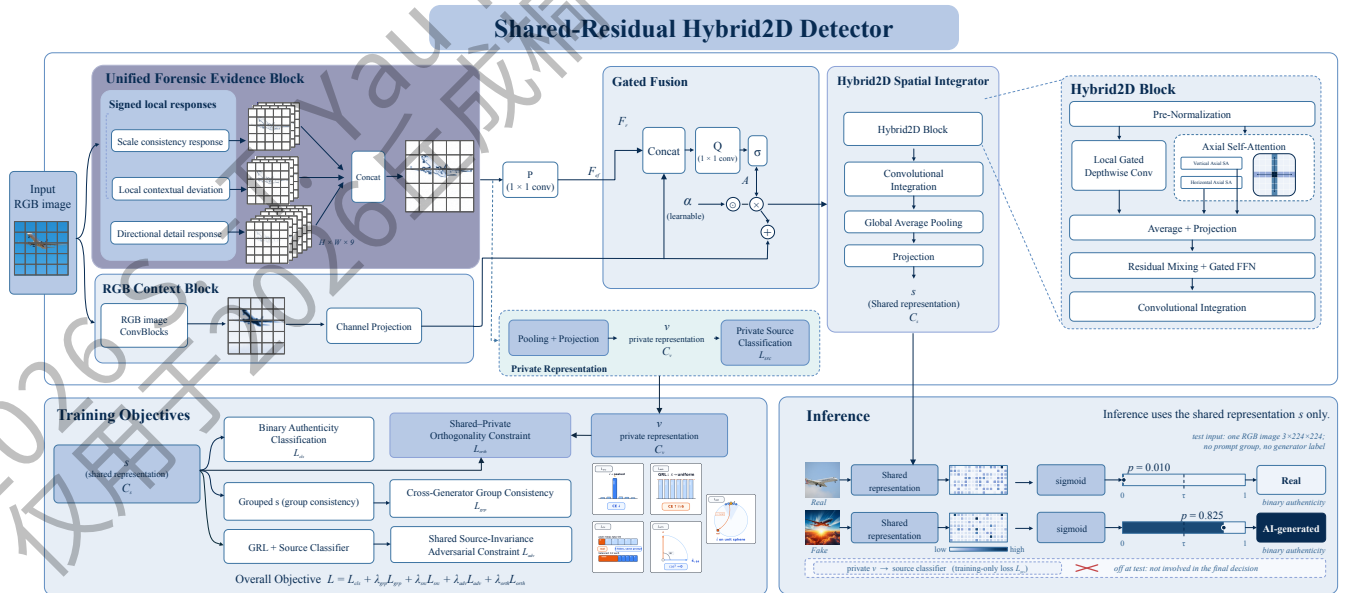


Figure 2. Framework overview. RGB and UFEB features undergo gated residual fusion and Hybrid2D integration for shared prediction; the pre-fusion UFEB features provide the private representation.

3 Method

The proposed framework consists of local forensic signal construction, content–forensic dual-branch fusion, and high-level spatial integration. As shown in Fig. 2, UFEB converts the raw image into complementary local responses, which are encoded separately from the RGB observed signal, fused through gated residual injection, and then integrated by Hybrid2D over local and cross-region dependencies. During training, within-group consistency and shared–private representation constraints organize the discriminative information; at inference, the shared representation performs authenticity classification.

Formally, the detector is a function that maps an image $\mathbf{x} \in [0, 1]^{3 \times H \times W}$ to a generation probability $p \in [0, 1]$, and the quantity of interest is its behavior on images whose generator was not seen during training. Let $\mathcal{G}_{\text{tr}}$ denote the visible generator set used for training and $\mathcal{G}_{\text{te}}$ the held-out generators; every fold trains on images from $\mathcal{G}_{\text{tr}}$ and tests on images from a generator outside it (Sec. 4.1). Two views of the same observed image are constructed at the input side. The forensic view is derived from the raw pixel values: fixed operators measure a multi-channel local response map $\mathbf{R}$ that describes how the image deviates from its own local structure (Sec. 3.1). The content view keeps the complete observed signal, including color, layout, and texture, after per-channel ImageNet normalization (Sec. 3.2). Each view is encoded by its own convolutional stages before the two meet, so both are built from the same pixels through different measurements of them.

The interaction between the two views is deliberately asymmetric. The content representation defines the output path, and forensic evidence is injected into it through a learnable gate (Sec. 3.3), so the decision stays grounded in the observed image while local evidence is added where it is informative. The fused map is then integrated over local and cross-region dependencies (Sec. 3.4), and two further representations, one shared and one private, are read from different points of the network and organized by the training objective (Sec. 3.5). The notation, the complete pipeline, and the parameter budget are summarized in Sec. 3.6. At training time the model also consumes prompt groups and generator labels; at inference it requires a single RGB image and no metadata.

3.1 Unified Forensic Evidence Block

UFEB derives three complementary classes of local image-domain response from the raw image. All operators of the block are fixed, so UFEB itself introduces no trainable parameters, and the three classes can be inspected separately from the learned encoder that follows. The scale response measures the multi-scale consistency error before and after downsampling and reconstruction, the contextual response quantifies the deviation of the center pixel relative to neighborhood statistics, and the directional response extracts oriented high-frequency variation. For an RGB image $\mathbf{x} \in [0, 1]^{3 \times H \times W}$, the three responses are defined as

$$\begin{aligned} \mathbf{R}_c^s &= \mathbf{x}_c - \mathcal{U}_2(\mathcal{D}_2(\mathbf{x}_c)), \\ R_c^c(u) &= x_c(u) - \text{ord}_4\{x_c(v) : v \in \mathcal{N}(u)\}, \\ \mathbf{R}_{c,k}^d &= \mathbf{K}_k * \mathbf{x}_c, \quad k = 1, 2, 3, \end{aligned} \tag{1}$$

where $\mathcal{D}_2$ and $\mathcal{U}_2$ halve and restore the spatial size by nearest-neighbor interpolation; $\mathcal{N}(u)$ is the 3×3 eight-neighborhood excluding the center, and ord_4 takes the fourth-smallest value. Both the neighborhood operations and the convolutions use reflection padding, so image borders do not produce artificial boundary responses. The three families describe different properties of the same local signal: the scale response is the residual between a pixel and its coarse-to-fine reconstruction, the contextual response compares a pixel with an order statistic that never contains the pixel itself and therefore tolerates a few extreme neighbors, and the directional response reacts only to oriented variation. All three are computed by fixed operators, so the evidence measured on an image does not depend on training and cannot drift with the learned parameters.

Three properties of the operators are worth stating, because they decide what the responses can represent. The scale response uses nearest-neighbor down- and upsampling, so its reference is assembled from the pixels that survive subsampling and the residual measures the deviation between a pixel and its coarsely represented value

without interpolation smoothing. The contextual reference is a lower median: up to three extreme values on either side of the neighborhood leave it unchanged, so a few very dark or very bright pixels cannot pull the reference away from the bulk of the neighborhood, and excluding the center keeps a pixel from masking its own deviation. Both responses are signed and are stored as such, so an overshoot and an undershoot at the same location remain different measurements. The directional responses are computed directly on the intensities without preceding smoothing, because the variation they target lies close to the sampling limit and would not survive blurring.

Directional detail is extracted by fixed high-pass kernels with zero DC response. Let $\mathbf{d} = (1, -2, 1)^\top$, $\mathbf{e} = (0, 1, 0)^\top$, and $\mathbf{t} = (-1, 0, 1)^\top$; the three kernels are

$$\begin{aligned}\mathbf{K}_1 &= -\frac{\mathbf{d}\mathbf{e}^\top + \mathbf{e}\mathbf{d}^\top}{8}, \\ \mathbf{K}_2 &= -\frac{\mathbf{d}\mathbf{d}^\top}{16}, \quad \mathbf{K}_3 = -\frac{\mathbf{t}\mathbf{t}^\top}{4}.\end{aligned}\tag{2}$$

Each kernel has unit sum of absolute values, so responses of different orders are comparable in magnitude, and the three kernels measure distinct spatial orders of local detail: $\mathbf{K}_1$ is a discrete Laplacian that reacts to isotropic curvature, $\mathbf{K}_2$ composes the second differences along the two axes and reacts to fourth-order variation, and $\mathbf{K}_3$ is the mixed second difference of the two first differences, i.e., a saddle detector. Every kernel is applied to the three color channels independently, so the directional class contributes nine signed maps; together with three scale and three contextual maps, the concatenated response $\mathbf{R}$ has 15 channels. No per-image normalization is applied, so the sign and the relative magnitude of every response reach the encoder unchanged, and responses from different images remain directly comparable to each other. The responses are computed on raw intensities in $[0, 1]$, and no fixed response is rescaled anywhere before the encoder, so the first normalization on the forensic side acts on the encoder outputs described next.

The response maps are encoded before they meet the RGB features. The response families mix isotropic residuals with oriented detail, so the encoder does not rely on one receptive-field shape: three parallel convolutions with kernels 3×5 , 5×5 , and 5×3 , each with stride 2, encode the response map,

$$\begin{aligned}\mathbf{E}^{(j)} &= \text{GELU}\left(\text{Norm}\left(\mathbf{K}^{(j)} * \mathbf{R}\right)\right), \quad j = 1, 2, 3, \\ \mathbf{F}_f^{(0)} &= \mathbf{W}_{\text{mix}}\left[\mathbf{E}^{(1)}, \mathbf{E}^{(2)}, \mathbf{E}^{(3)}\right],\end{aligned}\tag{3}$$

where $\mathbf{K}^{(j)}$ is the j -th kernel, $\mathbf{W}_{\text{mix}}$ is a 1×1 convolution, and the three branches split the stem width into three nearly equal groups (10, 10, and 12 of the 32 stem channels). Each branch normalizes and activates its own output before the three are recombined, so every receptive-field shape contributes its own channel group to the mixed map. The raw responses enter the first convolution without pre-normalization, and normalization acts on the convolution outputs instead, so the signed evidence itself is never rescaled. Stride 2 matches the RGB stem (Sec. 3.2), so the two branches keep the same resolution and channel width. The resulting map $\mathbf{F}_f^{(0)}$ is the first forensic feature map, on which the residual stages of the forensic branch operate before the two branches are fused. The three response families can also be removed one at a time from this construction, which is how their individual contribution is measured in Sec. 4.6 (Table 6).

3.2 RGB Block

The RGB block keeps the complete observation of the image. The responses of UFEB are deliberately narrow: they preserve local residuals and discard color, layout, and untouched content. The RGB branch is the complement of that view, and it remains the stream that carries evidence when the local responses are weak or dispersed. Its input is the ImageNet-normalized image $\mathbf{I}$, i.e., the image normalized per channel with the ImageNet mean and standard deviation, and no fixed operator is applied to the pixels before learning. A single strided convolution forms the stem,

$$\mathbf{F}^{(0)} = \text{GELU}(\text{Norm}(\mathbf{K}_{7 \times 7} * \mathbf{I})),\tag{4}$$

with kernel size 7×7 and stride 2. The large kernel covers a broad neighborhood in one layer, which suits the color and texture statistics that are already visible at this scale, and the stride halves the resolution so that the later stages operate on feature maps instead of pixels. The forensic stream uses the same stride in its encoder (Sec. 3.1), so the two branches stay spatially aligned: for 224×224 inputs both produce a 32-channel map at 112×112 , and the whole network keeps this common sampling grid.

The stem and both shallow stages share one block design. The block applies two 3×3 convolutions with pre-normalization, i.e., a normalization before each of them, and a GELU in between, and it reweights the channels with the efficient channel attention (ECA) module⁴³ before the scaled residual addition:

$$\begin{aligned} \mathbf{Y} &= \mathbf{W}_2 \text{Norm}(\text{GELU}(\mathbf{W}_1 \text{Norm}(\mathbf{X}))), \\ \mathbf{X}' &= \mathbf{X} + \gamma \odot \text{ECA}(\mathbf{Y}), \end{aligned} \quad (5)$$

$$\text{ECA}(\mathbf{Y}) = \mathbf{Y} \odot \sigma(\mathbf{W}_c \text{GAP}(\mathbf{Y})), \quad (6)$$

where $\mathbf{W}_1, \mathbf{W}_2$ are 3×3 convolutions, GAP returns one average per channel, $\mathbf{W}_c$ is a one-dimensional convolution with kernel 3 along the channel axis, and γ is a per-channel scale initialized at 10^{-5} , so every block starts close to an identity mapping and the convolutions of a block are introduced gradually during training. The attention is driven by the descriptor of the feature map that it modulates, and it adds three parameters per block, which keeps the channel reweighting inexpensive. The two 3×3 convolutions keep the width unchanged while enlarging the receptive field, and because the reweighted branch output is scaled by γ before it is added, the identity path dominates at the beginning of training and each block behaves like a shallow feature extractor until its parameters have moved. The same blocks process the forensic stream, so the two branches differ in what they observe rather than in how their stages are built.

Between stages, the resolution is halved and the width doubled by the anti-aliased downsampling used throughout the network⁴⁴,

$$\mathbf{F}_d = \text{LP}_{[1,2,1]}(\mathbf{W}_d \mathbf{F}), \quad (7)$$

where $\mathbf{W}_d$ is a 1×1 channel projection and $\text{LP}_{[1,2,1]}$ is a separable $[1, 2, 1]$ low-pass filter applied per channel with stride 2 and reflection padding. Filtering before subsampling keeps the operation from folding high-frequency content back into the feature map, which matters here because part of the evidence for and against an image lies in high-frequency structure. Each downsampling step doubles the width while halving the resolution, so the capacity of the representation grows as its spatial detail is compressed. In the reported configuration, one block follows the stem and one follows the first downsampling stage, with widths of 32 and 64 channels; the second downsampling stage doubles the width to 128, where the two streams meet. The block ends with the content representation $\mathbf{F}_r$, i.e., the channel projection taken at the fusion width and resolution; for 224×224 inputs this is a 128-channel map at 28×28 .

3.3 Gated Residual Fusion

The two branches are encoded independently at shallow stages and meet only here: their outputs, the content representation $\mathbf{F}_r$ and the forensic representation $\mathbf{F}_f$, share the same resolution and width. Forensic information is injected into the RGB representation through a gated residual:

$$\begin{aligned} \tilde{\mathbf{F}}_f &= \mathcal{P}(\mathbf{F}_f), \\ \mathbf{A} &= \sigma\left(\mathcal{Q}([\mathbf{F}_r, \tilde{\mathbf{F}}_f])\right), \\ \mathbf{F} &= \mathbf{F}_r + \alpha \mathbf{A} \odot \tilde{\mathbf{F}}_f, \end{aligned} \quad (8)$$

where the projection $\mathcal{P}$ and the gating mapping $\mathcal{Q}$ are both 1×1 convolutions, $\mathbf{A}$ modulates the injection strength at each spatial location and channel, and α is a learnable scalar initialized at 1. The gating mapping is initialized

with zero weights and a bias of -2 , so every gate value starts at $\sigma(-2) \approx 0.12$: the injection is weak and uniform at the beginning of training, and the gate has to move away from its initial value before forensic evidence can influence the decision. The gate reads both representations but only the projected forensic map is injected, so the content path passes through the fusion layer without being transformed. The RGB identity path is preserved throughout fusion, so the forensic stream cannot overwrite the content representation before the gate has learned to rely on it. At the fusion width of 128 channels the two 1×1 convolutions hold about 5×10^4 weights together, roughly 3% of the model (Sec. 3.6). After this step the dual-branch part of the network ends: the fused map continues into the integration stages, while the forensic map is set aside for the private representation.

3.4 Hybrid2D Spatial Integration

The fused map carries two kinds of structure at the same time. Many generation traces are short-range deviations, for example the local color and brightness dependencies left by an upsampling step, but whether such a deviation is a trace or ordinary image content usually cannot be decided from the deviation alone; it depends on the surrounding region. A stack of local convolutions captures short-range structure efficiently, but distant regions can only be related after many layers. Dense two-dimensional attention relates them in a single step, at a cost that grows quadratically with the number of pixels. Hybrid2D keeps the two computations apart inside one block and merges their outputs afterwards, and it is used only at the two deepest stages, where the maps are small enough for sequence mixing to be affordable (Sec. 3.6).

Let $\hat{\mathbf{F}} = \text{Norm}(\mathbf{F})$ be the pre-normalized block input. The local path expands the channel dimension by a factor of $3/2$, applies a 3×3 depthwise convolution, and splits the result into a content term and a gate,

$$\begin{aligned} [\mathbf{U}_c, \mathbf{U}_g] &= \text{DW}_{3 \times 3}(\mathbf{W}_{\text{in}} \hat{\mathbf{F}}), \\ \mathbf{Y}_{\text{loc}} &= \mathbf{W}_{\text{out}}(\text{SiLU}(\mathbf{U}_c) \odot \sigma(\mathbf{U}_g)), \end{aligned} \quad (9)$$

where $\mathbf{W}_{\text{in}}$ and $\mathbf{W}_{\text{out}}$ are 1×1 convolutions. The gate is applied per channel and per spatial position, so the mixer can attenuate locations where the depthwise response is weak. At the fusion width the expansion turns 128 channels into a hidden width of 192, and a dropout of 0.1 is applied to the gated output before it is projected back to the block width.

The second path applies axial self-attention⁴⁵ along the two spatial axes separately. Within one column the feature vectors form a sequence of length H , and within one row they form a sequence of length W ; each direction has its own projections and uses four heads of dimension $d_h = C/4$ (i.e., 32 dimensions per head at the fusion width),

$$\begin{aligned} \mathbf{Z}^a &= \text{softmax} \left(\frac{\mathbf{Q}^a (\mathbf{K}^a)^\top}{\sqrt{d_h}} \right) \mathbf{V}^a, \quad a \in \{x, y\}, \\ \mathbf{Y}_{\text{ax}} &= \mathbf{W}_o \left(\frac{1}{2} (\mathbf{Z}^x + \mathbf{Z}^y) \right), \end{aligned} \quad (10)$$

where $a = x$ refers to the row direction and $a = y$ to the column direction, and $(\mathbf{Q}^a, \mathbf{K}^a, \mathbf{V}^a)$ are obtained from the corresponding sequence by the projections of that direction. Because both attentions are one-dimensional, the cost is $\mathcal{O}(HW(H+W))$ rather than $\mathcal{O}((HW)^2)$, which keeps the block affordable at the fusion resolution. The outputs of the two directions are averaged before projection, so that both directions contribute to every position, and a dropout of 0.1 is applied to the projected output. No positional embedding is added to either direction: the position-dependent structure of the block is supplied by the depthwise local path, and the attention relates positions through their content. At the fusion resolution the sequences have length 28, so the two axial directions together cost 43,904 attention entries, an order of magnitude below the 614,656 entries required by dense attention over the same map, which keeps the two attention stages affordable at this parameter scale.

The two paths are added and used to update the block input, and a gated feed-forward subnetwork follows:

$$\begin{aligned} \mathbf{F}' &= \mathbf{F} + \gamma_1 \odot (\mathbf{Y}_{\text{loc}} + \mathbf{Y}_{\text{ax}}), \\ \mathbf{F}'' &= \mathbf{F}' + \gamma_2 \odot \mathcal{M}_{\text{ffn}}(\text{Norm}(\mathbf{F}')), \end{aligned} \quad (11)$$

$$\mathcal{M}_{\text{fin}}(\mathbf{X}) = \mathbf{W}_{f2}(\text{SiLU}(\mathbf{H}_2) \odot \mathbf{H}_1), \quad [\mathbf{H}_1, \mathbf{H}_2] = \text{DW}_{3 \times 3}(\mathbf{W}_{f1} \mathbf{X}), \quad (12)$$

where $\gamma_1, \gamma_2 \in \mathbb{R}^C$ are per-channel scales initialized at 10^{-5} , so every block starts close to an identity mapping and the two residual branches enter gradually. The gated feed-forward subnetwork follows the SwiGLU pattern: a pointwise projection produces two hidden halves, a depthwise convolution mixes each half locally, the activation of the second half multiplies the first, and a final pointwise projection returns to the block width. The hidden width of the feed-forward subnetwork is twice the block width. Stochastic depth is applied to both branches, with rates that increase linearly with block depth up to 0.15.

After fusion, two Hybrid2D blocks operate at the fusion resolution, where the response maps are still fine enough for row- and column-wise dependencies to be informative; together with the blocks of the shallow stages, the four stages of the network contain one, one, two, and one blocks, respectively. For 224×224 inputs the four stages operate at 112×112 , 56×56 , 28×28 , and 14×14 with widths 32, 64, 128, and 256. A convolutional integration stage then follows: a downsampling step of the same construction as in Sec. 3.2 halves the spatial size and doubles the channel width, and a last block, which keeps the local path alone, integrates the result at the reduced resolution; the last stage drops the attention path because at the smallest resolution the local window already relates a larger fraction of the map and the row- and column-wise relations have been mixed at the fusion resolution. A normalization with global average pooling reduces the output of this stage to a vector, and a linear layer followed by layer normalization and a GELU activation forms the shared representation $\mathbf{s}$, which is the only quantity that the classification head reads at inference. The pre-fusion forensic representation $\mathbf{F}_f$ bypasses the integration path; it is pooled and projected by a separate head into the private representation $\mathbf{v}$, which carries the source-related supervision used during training. Because the two vectors are read from different points of the network, the integration stages cannot change what the private representation contains, and the only training signals that reach the shared representation from the source side are the adversarial and orthogonality terms (Sec. 3.5). In the reported configuration, the shared representation has 128 dimensions and the private representation has 64.

3.5 Training Objective and Inference

The training objective combines five terms: a class-balanced binary classification loss $\mathcal{L}_{\text{cls}}$, a within-group consistency loss $\mathcal{L}_{\text{grp}}$, a private source-classification loss $\mathcal{L}_{\text{src}}$, an adversarial source-confusion loss $\mathcal{L}_{\text{adv}}$, and an orthogonality loss $\mathcal{L}_{\text{orth}}$. Each term is described separately below.

Classification loss. The primary term is a class-balanced binary cross-entropy on the shared representation. A single linear head maps $\mathbf{s}$ to a logit, and the two classes receive equal total weight,

$$\mathcal{L}_{\text{cls}} = \sum_{i \in \mathcal{B}} w_i \ell_{\text{bce}}(\mathbf{w}_c^\top \mathbf{s}_i + b_c, y_i), \quad (13)$$

$$w_i = \begin{cases} 0.5/n_0, & y_i = 0, \\ 0.5/n_1, & y_i = 1, \end{cases}$$

where $y_i \in \{0, 1\}$ labels the real and synthetic samples, n_0 and n_1 count the two classes in the batch, and ℓ_{bce} is the binary cross-entropy with logits. In the reported setting a batch holds twelve images, two real ones and ten synthetic ones, so each real image receives weight 0.25 and each synthetic image 0.05: the two classes contribute 0.5 each, and the real image of a group is not outvoted five to one.

Group consistency loss. To encourage cross-generator consistency, each training group q contains one real image together with images from the $M = |\mathcal{G}_{\text{tr}}|$ visible generators within the same prompt group, where $\mathcal{G}_{\text{tr}}$ excludes the held-out generator. With a bar denoting ℓ_2 normalization, μ_q is the normalized mean of the within-group $\bar{\mathbf{s}}_{q,g}$, i.e., the group centroid. The constraint is

$$\mathcal{L}_{\text{grp}} = \frac{1}{|\mathcal{B}_{\text{grp}}|} \sum_{q \in \mathcal{B}_{\text{grp}}} \left[\frac{1}{M} \sum_{g \in \mathcal{G}_{\text{tr}}} \left(1 - \bar{\mathbf{s}}_{q,g}^\top \mu_q \right) + \eta \left[m - 1 + \bar{\mathbf{s}}_{q,0}^\top \mu_q \right]_+ \right], \quad (14)$$

where $\mathcal{B}_{\text{grp}}$ is the set of groups in a batch, the subscript 0 denotes the real sample, m is the separation margin, and η weights the separation term. The first part is the mean cosine distance between the M synthetic embeddings and their centroid, so the generator embeddings of the same prompt are pulled towards one common direction; the second part is a hinge on the real sample. The hinge vanishes while the real embedding stays farther than cosine distance m from the centroid and grows linearly once it falls below that distance, so m sets the separation required between the real image and the synthetic members of its group, and η trades this separation against the within-group alignment. The loss is averaged over the groups in the batch and requires only semantic correspondence between group members, not pixel-level registration. The reported configuration uses $m = 0.5$ and $\eta = 0.25$ (Sec. 4.1).

The shared-private part of the objective separates common discriminative information from source-dependent information.

Private source-classification loss. The private branch is trained to recognize the generator of a synthetic image. For every synthetic sample in the batch, a single linear classifier reads $\mathbf{v}$ and predicts one of the visible generators, and the loss is the mean softmax cross-entropy over those samples,

$$\begin{aligned}\mathcal{L}_{\text{src}} &= \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \ell_{\text{ce}}(\mathbf{W}_{\text{p}} \mathbf{v}_i + \mathbf{b}_{\text{p}}, g_i), \\ \mathcal{L}_{\text{adv}} &= \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \ell_{\text{ce}}(\mathbf{W}_{\text{a}} \mathcal{R}_1(\mathbf{s}_i) + \mathbf{b}_{\text{a}}, g_i),\end{aligned}\tag{15}$$

where $\mathcal{S} \subseteq \mathcal{B}$ is the set of synthetic samples, g_i is the generator label of sample i , ℓ_{ce} is the softmax cross-entropy, and the two classifiers are independent linear layers over $\mathbf{v}$ and over $\mathbf{s}$, respectively. Minimizing $\mathcal{L}_{\text{src}}$ keeps the information that identifies the generator inside the private representation.

Adversarial source-confusion loss. The second line of Eq. 15 turns the same supervision against the shared representation. The operator $\mathcal{R}_1$ is a gradient-reversal layer³⁹: it returns its input unchanged in the forward pass and multiplies the arriving gradient by -1 in the backward pass. The shared-side classifier therefore still learns to identify generators, but the gradient that it sends back to $\mathbf{s}$ maximizes its own cross-entropy instead of minimizing it, so the shared representation is pushed towards a source-uniform distribution while the private representation keeps separating the generators. Both classifiers are used only during training, and the reversal coefficient of $\mathcal{R}_1$ is 1, i.e., the gradient is inverted at full strength.

Orthogonality loss. The last term keeps the private information from being mirrored inside the shared representation. A fixed coordinate selection matrix $\mathbf{P}$ extracts a prefix of the shared representation with the same length as the private representation, i.e., the first 64 coordinates of $\mathbf{s}$ in the reported configuration, and the loss is the mean squared cosine between the normalized prefix and the normalized private vector,

$$\mathcal{L}_{\text{orth}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left(\overline{\mathbf{P}} \mathbf{s}_i^\top \bar{\mathbf{v}}_i \right)^2,\tag{16}$$

where $\mathcal{B}$ denotes all images in the batch and the two vectors are ℓ_2 -normalized, so each summand lies in $[0, 1]$: it is zero when the prefix and the private vector are orthogonal and one when they are parallel. The constraint applies to the part of the shared representation that has the same width as the private representation, so the source-related information carried by $\mathbf{v}$ cannot simply be re-read from that part of $\mathbf{s}$.

Overall objective and inference. The five terms are combined with fixed weights,

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{grp}} \mathcal{L}_{\text{grp}} + \lambda_{\text{src}} \mathcal{L}_{\text{src}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{orth}} \mathcal{L}_{\text{orth}},\tag{17}$$

where each λ controls the weight of the corresponding auxiliary objective, and the reported configuration weights the four auxiliary terms as 0.20, 0.10, 0.10, and 0.01 (Sec. 4.1), so the binary decision remains the primary task while the four auxiliary terms shape the two representations. At inference the shared representation alone is used: the classification head of Eq. 13 outputs the generation probability $p = \sigma(\mathbf{w}_{\text{c}}^\top \mathbf{s} + b_{\text{c}})$, requiring neither prompts nor

source labels; the private representation, both generator classifiers, and the group construction take no part in the decision. A binary verdict is obtained by comparing p with a threshold τ (Fig. 2), whereas the evaluation metrics reported below are threshold-free.

3.6 Method Summary

This section closes the method description with the same material in compact form. Table 2 collects the notation used in Sec. 3, grouped by pipeline stage, and Algorithm 1 restates the pipeline end to end, from the pixel input to the decision and the training step, annotating each step with its defining equation. The closing paragraph reports how the trainable parameters of the reported configuration are distributed over the components of the model.

Table 2. Notation used in Sec. 3, grouped by pipeline stage. Symbols confined to a single equation are defined at first use.

Notation	Meaning
<i>Input and forensic evidence (Sec. 3.1)</i>	
$\mathbf{x} \in [0, 1]^{3 \times H \times W}; \mathbf{I}$	raw input image; its ImageNet-normalized copy (content view)
$\mathbf{R}; \mathbf{R}^s, \mathbf{R}^c, \mathbf{R}^d$	concatenated signed response map (15 channels); scale, contextual, and directional response families (Eq. 1)
$\mathbf{K}_1, \mathbf{K}_2, \mathbf{K}_3; \mathbf{K}^{(j)}; \mathbf{E}^{(j)}$	fixed zero-DC detail kernels (Eq. 2); encoder kernels of shape $3 \times 5, 5 \times 5, 5 \times 3$; encoded responses (Eq. 3)
<i>Content encoding and fusion (Secs. 3.2 and 3.3)</i>	
$\mathbf{F}^{(0)}, \mathbf{F}_f; \mathbf{F}_f^{(0)}, \mathbf{F}_f$	RGB stem map (Eq. 4) and pre-fusion content representation (128 channels at 28×28); first forensic map and pre-fusion forensic representation
$\text{ECA}(\cdot); \gamma; \text{LP}_{[1,2,1]}(\cdot)$	channel attention (Eq. 6); per-channel residual scale (initialized at 10^{-5}); anti-aliased blur-pool downsampling (Eq. 7)
$\mathcal{P}, \mathcal{Q}; \mathbf{A}, \alpha; \mathbf{F}$	fusion projections (1×1 convolutions); gate map and injection scale (initialized at 1); fused representation after gated injection (Eq. 8)
<i>Integration and readout (Sec. 3.4)</i>	
$\mathbf{Y}_{\text{loc}}, \mathbf{Y}_{\text{ax}}; \gamma_1, \gamma_2; \mathcal{M}_{\text{ffn}}$	local-mixer and axial-attention outputs; per-channel block scales (initialized at 10^{-5}); gated feed-forward subnetwork (Eqs. 9–12)
$d_h; \mathbf{Q}^a, \mathbf{K}^a, \mathbf{V}^a; \mathbf{W}_o$	attention head dimension $C/4$; per-axis queries, keys, and values; axial output projection (Eq. 10)
$\mathbf{s} \in \mathbb{R}^{128}; \mathbf{v} \in \mathbb{R}^{64}$	shared representation (decision path); private representation (training only)
<i>Training objective (Sec. 3.5)</i>	
$y_i; g_i; n_0, n_1; w_i$	real–synthetic label; generator label of synthetic sample i ; class counts; class-balanced weights (Eq. 13)
$\ell_{\text{bce}}, \ell_{\text{ce}}; \mu_q; m, \eta$	binary and softmax cross-entropy; normalized group centroid; separation margin and its weight (Eqs. 13 and 14)
$M; \mathcal{B}, \mathcal{B}_{\text{grp}}; \mathcal{I}$	number of visible generators; batch and its prompt groups; synthetic samples in the batch
$\mathcal{L}_{\text{cls}}, \mathcal{L}_{\text{grp}}; \mathcal{L}_{\text{src}}, \mathcal{L}_{\text{adv}}, \mathcal{L}_{\text{orth}}$	class-balanced classification and group consistency losses; private source, adversarial source-confusion (through the gradient reversal $\mathcal{R}_1$), and orthogonality losses (Eqs. 13–16)
$\mathbf{P}; \lambda_{\text{grp}}, \lambda_{\text{src}}, \lambda_{\text{adv}}, \lambda_{\text{orth}}; p, \tau$	prefix selection (first 64 coordinates of $\mathbf{s}$); auxiliary weights (0.20, 0.10, 0.10, 0.01); generation probability and binary decision threshold (Eqs. 16 and 17)

Algorithm 1 UFEB–Hybrid2D detector: dual-branch forward pass and training step.

Require: image $\mathbf{x}$; labels $\{y_i, g_i\}$ during training; threshold τ at inference

```
1:  $\mathbf{R} \leftarrow \text{UFEB}(\mathbf{x})$  ▷ forensic view: 15 signed response maps, fixed operators (Eq. 1)
2:  $\mathbf{I} \leftarrow \text{ImageNetNorm}(\mathbf{x})$  ▷ content view
3:  $\mathbf{E}^{(j)} \leftarrow \text{GELU}(\text{Norm}(\mathbf{K}^{(j)} * \mathbf{R}))$ ,  $j = 1, 2, 3$ 
4:  $\mathbf{F}_f^{(0)} \leftarrow \mathbf{W}_{\text{mix}}[\mathbf{E}^{(1)}, \mathbf{E}^{(2)}, \mathbf{E}^{(3)}]$  ▷ forensic branch: encoder of Eq. 3
5:  $\mathbf{F}^{(0)} \leftarrow \text{GELU}(\text{Norm}(\mathbf{K}_{7 \times 7} * \mathbf{I}))$  ▷ RGB branch: stem of Eq. 4
6: for  $k = 1, 2$  do
7:    $\mathbf{F}_f^{(k)} \leftarrow \text{LP}_{[1,2,1]}(\mathbf{W}_d \text{Block}(\mathbf{F}_f^{(k-1)}))$  ▷ forensic branch, stage  $k$  (Eqs. 5 and 7)
8:    $\mathbf{F}^{(k)} \leftarrow \text{LP}_{[1,2,1]}(\mathbf{W}_d \text{Block}(\mathbf{F}^{(k-1)}))$  ▷ RGB branch, stage  $k$ : same block and sampling grid
9: end for
10:  $\mathbf{F}_f, \mathbf{F}_r \leftarrow \text{fusion-width projections (128 channels at } 28 \times 28)$ 
11:  $\mathbf{A} \leftarrow \sigma(\mathcal{Q}([\mathbf{F}_r, \mathcal{P}(\mathbf{F}_f)]))$ ;  $\mathbf{F} \leftarrow \mathbf{F}_r + \alpha \mathbf{A} \odot \mathcal{P}(\mathbf{F}_f)$  ▷ branches meet: gated injection of Eq. 8
12: for  $l = 1, 2$  do
13:    $\mathbf{F} \leftarrow \mathbf{F} + \gamma_1 \odot (\mathbf{Y}_{\text{loc}} + \mathbf{Y}_{\text{ax}})$ ;  $\mathbf{F} \leftarrow \mathbf{F} + \gamma_2 \odot \mathcal{M}_{\text{ffn}}(\text{Norm}(\mathbf{F}))$  ▷ single stream: Hybrid2D block (Eqs. 9–12)
14: end for
15:  $\mathbf{F} \leftarrow \text{Block}_{\text{local}}(\text{LP}_{[1,2,1]}(\mathbf{W}_d \mathbf{F}))$  ▷ last stage: downsample, local path only
16:  $\mathbf{s} \leftarrow \text{GELU}(\text{LN}(\text{Linear}(\text{GAP}(\mathbf{F}))))$ ;  $\mathbf{v} \leftarrow \text{Linear}(\text{GAP}(\mathbf{F}_f))$  ▷ shared and private readouts
17:  $p \leftarrow \sigma(\mathbf{w}_c^\top \mathbf{s} + b_c)$ ; verdict  $\leftarrow \mathbf{1}[p \geq \tau]$  at inference
18: if training then
19:    $\mathcal{L} \leftarrow \mathcal{L}_{\text{cls}} + \lambda_{\text{grp}} \mathcal{L}_{\text{grp}} + \lambda_{\text{src}} \mathcal{L}_{\text{src}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{orth}} \mathcal{L}_{\text{orth}}$  ▷ Sec. 3.5
20:    $\theta \leftarrow \theta - \text{lr} \cdot \nabla_{\theta} \mathcal{L}$  ▷ source classifiers and  $\mathbf{v}$  are training-only
21: end if
22: return  $p$  (and  $\mathbf{s}$  during training)
```

The trainable budget of the reported configuration is 1.64M parameters (Sec. 4.1), distributed as follows. The fixed evidence bank contributes none of them, and channel attention adds three parameters per convolutional block, so the forensic measurement side is essentially free. The gated fusion layer holds two 1×1 convolutions with about 0.05M weights, roughly 3% of the model, and the four convolutional blocks of the shallow stages add about 0.19M. The two integration stages hold close to four fifths of the budget: the two Hybrid2D blocks at the fusion resolution account for about 0.59M, and the single block of the last stage, which is twice as wide, accounts for about 0.71M by itself, so the capacity is concentrated where the fused representation is integrated. Every component scales linearly with the number of pixels except the axial attention of Sec. 3.4, whose per-direction cost is $HW(H + W)$, which keeps the cost profile of a lightweight image classifier. Dropout at rates 0.1, 0.1, and 0.2, stochastic depth up to 0.15, group normalization with 32 groups, and layer normalization on the two representation vectors complete the configuration.

4 Experiments

4.1 Experimental Setup

The experiments use the COCO_AI dataset¹, which contains 10,017 semantically corresponding groups, each consisting of one real COCO image and images produced by six generators. The groups are split into 7,011/1,001/2,005 training, validation, and test groups, corresponding to 42,066, 6,006, and 4,010 images per fold, and groups containing repeated prompts or images do not cross splits. We use six-fold Leave-One-Generator-Out (LOGO) evaluation: each fold trains and validates on the five non-held-out generators and real images, then tests on 2,005 images

¹https://huggingface.co/datasets/NasrinImp/COCO_AI

from the held-out generator and 2,005 real images. This setting evaluates transfer to an unseen generative signal distribution. Models are selected by validation AUROC on the visible generators, without tuning on the held-out generator.

All models are trained from scratch using AdamW with seed 42 and weight decay 10^{-4} . Learning rates are 10^{-4} for baselines and 3×10^{-4} for our model, which uses two prompt groups per batch; each group contains one real image and one image from every visible generator, so a batch holds twelve images in total. Images are center-cropped to squares and resized to 256×256 , followed by 224×224 random training crops or center test crops. Training augmentation includes horizontal flipping, JPEG compression, and Gaussian blur. We set $(\lambda_{\text{grp}}, \lambda_{\text{src}}, \lambda_{\text{adv}}, \lambda_{\text{orth}}) = (0.20, 0.10, 0.10, 0.01)$, $m = 0.5$, and $\eta = 0.25$. The metrics are defined as in Table 4; each fold’s TPR_1 is the highest test-ROC TPR at $\text{FPR} \leq 1\%$, and JPEG compression and Gaussian blur additionally serve as training perturbations for common image-signal distortions. Training stops early on the validation AUROC of the visible generators, with a minimum improvement of 5×10^{-4} .

4.2 Dataset Analysis

COCO_AI is organized in prompt groups rather than in independent images. Each group holds one COCO photograph and six synthetic images, one from every generator, all produced from the same caption, so the semantic content is fixed within a group and only the generation process varies. Fig. 3 shows a test group together with the prompt that generated it, “A giraffe walking in a field with gazelle’s looking on.”; the real photograph is landscape, in the aspect ratio of its source image, while the six generator images are square, and they agree on the scene while differing in style, layout, and native resolution. This construction matters for evaluation, because the two classes cannot be separated by the objects they contain: a decision has to depend on the generation trace of the synthetic image rather than on the scene it depicts.



Figure 3. One test group of COCO_AI. The caption is shared by the real photograph and the six generator images; the generators agree on the scene and differ in style, background layout, and native resolution, from 270×270 (DALL-E3) and 436×436 (Midjourney6) to 768×768 and 1024×1024 for the Stable Diffusion models.

Counting all seven sources, the dataset contains 70,119 images in a one-to-six real-to-synthetic ratio (10,017 real against 60,102 synthetic), distributed over the three splits as 49,077, 7,007, and 14,035 images, with each source contributing 10,017 images (Fig. 4(a,b)). A LOGO fold removes the images of its held-out generator, so the number of images available to a model is six of the seven images per group, that is, the 42,066/6,006/4,010 figures of Sec. 4.1. The 2,005 real test images are moreover shared by all six folds (Fig. 4(c)), which makes the folds directly comparable with one another and attributes their differences to the held-out generator alone.

The two classes also differ in their native image properties (Table 3 and Fig. 5). The real photographs are heterogeneous, spanning 842 distinct resolutions over a short side between 128 and 640 pixels, with a landscape median of 640×480 , whereas the generator images are nearly uniform: each of the two smaller Stable Diffusion versions emits a single resolution (768×768 or 1024×1024) and DALL-E3 and Midjourney6 emit two. The encoded file sizes vary by almost an order of magnitude, from a median of 17.9 KiB for DALL-E3 to 140.8 KiB for SD3.5, and at the common 1024×1024 resolution the medians still grow from 111.8 KiB for SDXL to 135.5 KiB for SD3 and 140.8 KiB for SD3.5, consistent with different amounts of retained fine detail at equal pixel count. The

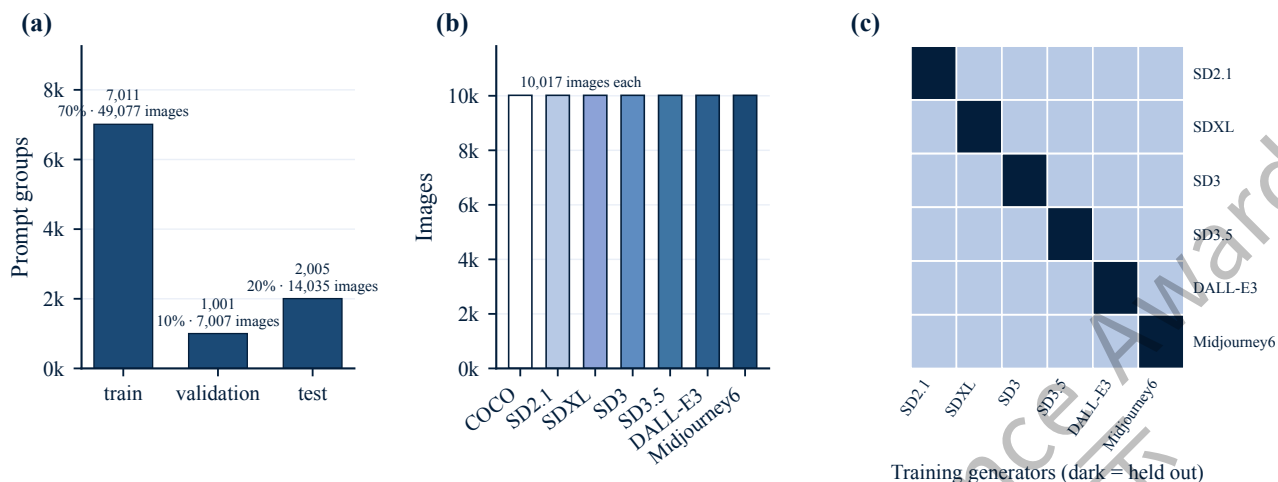


Figure 4. Dataset composition. (a) Prompt groups per split with the corresponding image counts; (b) images per source; (c) the six leave-one-generator-out folds, where the dark cell marks the held-out generator and the remaining five generators, together with the real images, form the training pool of the fold.

captions are short and concrete, with a median of ten words and between eight and thirteen words across the first to ninth decile (Fig. 5(d)).

Table 3. Per-source image properties of COCO_AI; each source contributes 10,017 images. Short side is $\min(w, h)$ per image, reported as median with the full range in parentheses; the resize ratio is 256 divided by the short side of the square crop, so a ratio of 1.0 means no resampling; the JPEG size is the median encoded file size over the dataset, which uses a single shared quantization table.

Source	Distinct res.	Short side (px)	Resize ratio	JPEG size (KiB)
COCO	842	428 (128–640)	0.60	48.6
SD2.1	1	768	0.33	88.3
SDXL	1	1024	0.25	111.8
SD3	1	1024	0.25	135.5
SD3.5	1	1024	0.25	140.8
DALL-E3	2	270 (270–351)	0.95	17.9
Midjourney6	2	436 (436–630)	0.59	24.5

Two potential shortcuts of the raw data were audited explicitly. First, every image is stored as JPEG with the same quantization table, so neither the container format nor the compression quality separates real from synthetic images. Second, the pixel dimensions are highly informative: on the stored files, a threshold on the short side alone reaches an AUROC of 1.000 on the four Stable Diffusion sources, 0.993 on DALL-E3, and 0.533 on Midjourney6, and only 2.9% of the real test images are square. The preprocessing chain removes this cue before the model sees an image, since every sample passes through the same center crop to a square and the same resize, which maps native short sides from 128 to 1024 pixels onto a common 224×224 input and leaves the two classes dimensionally indistinguishable. What the chain does not equalize is the resampling factor, which is 0.25 for the 1024×1024 images, 0.95 for DALL-E3, and close to 0.6 for the real photographs and Midjourney6 (Fig. 5(b)), so a common input format does not imply a common resampling history. With the format and dimension cues either absent or removed, the results of the following sections reflect the transferable generation traces of the six generators under this fixed chain rather than storage-level properties of the files.

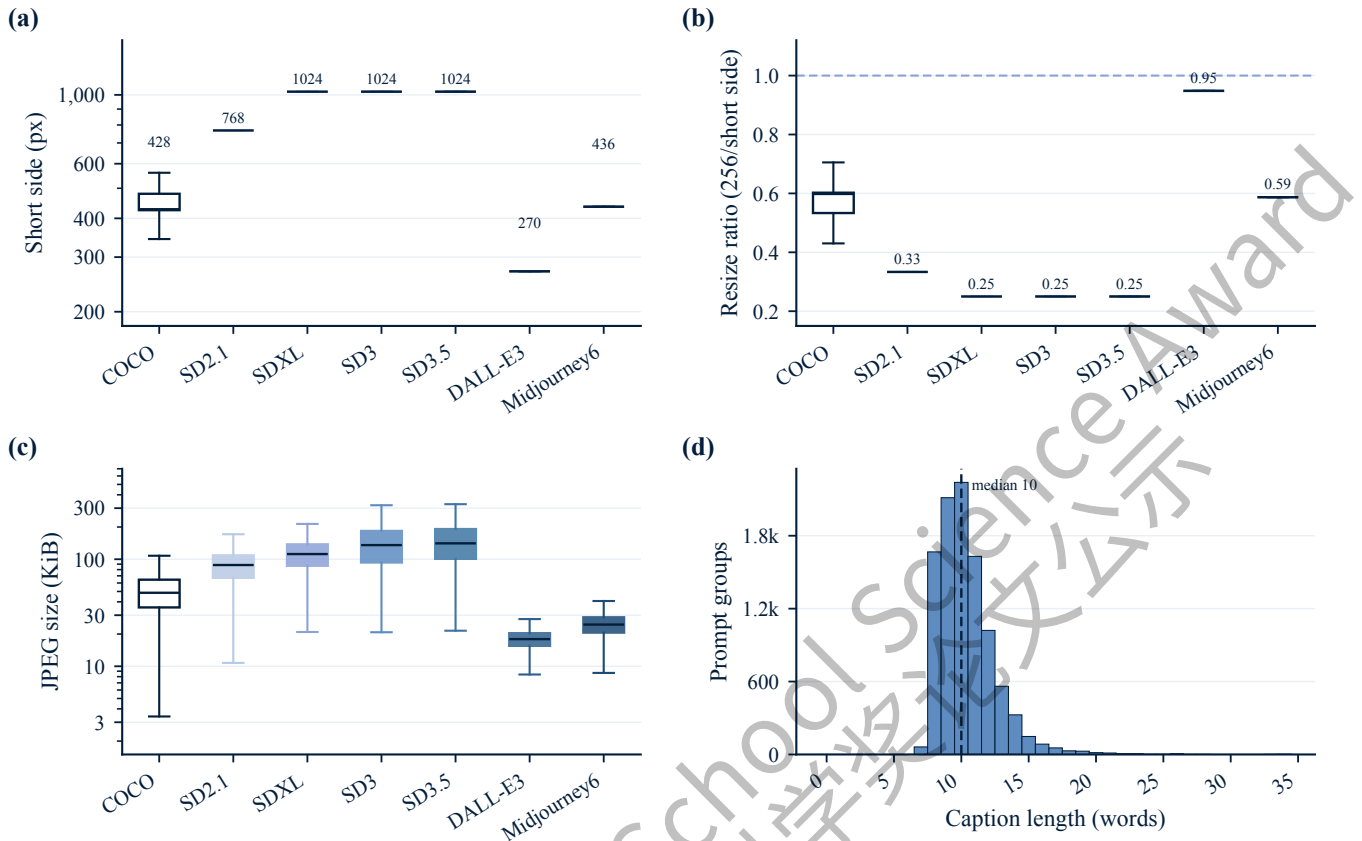


Figure 5. Image properties per source. (a) Native short side on a logarithmic scale; (b) resize ratio of the preprocessing chain, with the dashed line marking no resampling; (c) encoded JPEG size on a logarithmic scale; (d) caption length in words over the 10,017 groups.

4.3 Model Training

The training loss of Fig. 6 falls smoothly from 0.87 to 0.41 without plateaus or spikes, and the spread between folds stays within 0.03 throughout; the individual fold curves follow the same trajectory, so the optimization proceeds in the same way whichever generator is held out, and the loss is still decreasing at the end of the plotted range.

Balanced accuracy is evaluated at a fixed threshold of 0.5 and averages the true-positive and true-negative rates, so it cannot be raised by favoring the larger class. Fig. 7 shows it climbing from 0.83 to 0.97 before it fluctuates around a steady level; the spread between folds narrows from 0.036 to below 0.01, and the one fold that starts visibly lower catches up with the rest instead of diverging.

The AUROC of Fig. 8 is rank-based and threshold-free. It rises from 0.944, passes 0.99 while the training loss is still falling steeply, and then improves gradually; the checkpoints selected by this curve reach 99.70% on the visible generators (Sec. 4.5), and the spread between folds narrows to below 0.001. Since the training loss is still decreasing at the end of the same range while neither validation criterion turns downward, no overfitting signature appears within the observed range: the pattern of a validation metric that peaks and then deteriorates is absent, and the model keeps consolidating its ranking ability as training continues.

4.4 Cross-Generator Generalization

Table 4 shows that our method attains the highest mean AUROC of 95.63% with 1.64M parameters, exceeding the best competing method CNNSpot by 5.53 percentage points; the worst-source AUROC reaches 88.47%, which

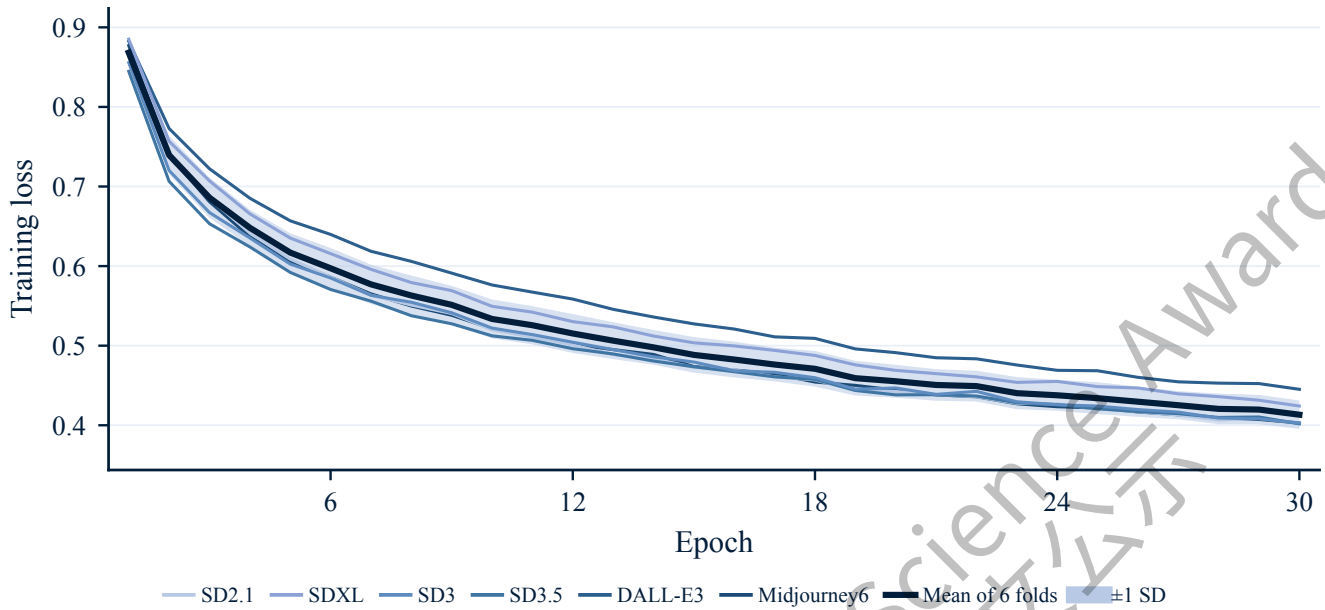


Figure 6. Training loss; the bold curve is the mean over the six leave-one-generator-out folds, the light curves are the individual folds, and the shaded band is the sample SD across folds.

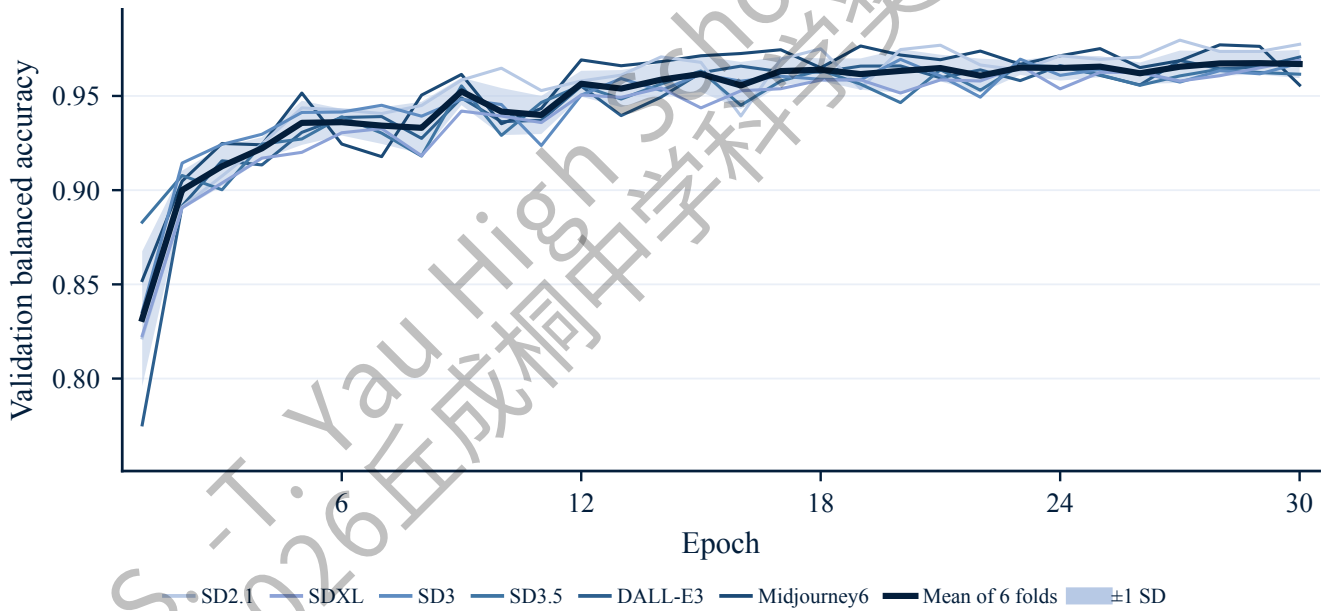


Figure 7. Validation balanced accuracy at a fixed threshold of 0.5; curves and band as in Fig. 6. Validation uses the five visible generators only.

is likewise the best result. The comparison spans detectors of very different sizes, from NPR and FerretNet with roughly one million parameters to a CLIP probe with 86.19M, and the accuracy-parameter view in Fig. 9 shows that the margin is not bought with capacity: CNNSpot needs 23.51M parameters, DINOv2 22.06M, CLIP 86.19M, and the ImageNet backbones ConvNeXt-T and Swin-T 27.82M and 27.52M to reach mean AUROCs of 90.11%, 81.77%, 75.85%, 79.38%, and 65.87%, whereas the two lightweight detectors remain either less accurate (FerretNet, 81.88% at 1.06M) or markedly less stable across sources (NPR, 90.03% at 1.44M with a sample SD of 13.25, against 4.56 for our model). NPR and FerretNet perform strongly on some Stable Diffusion sources, but degrade

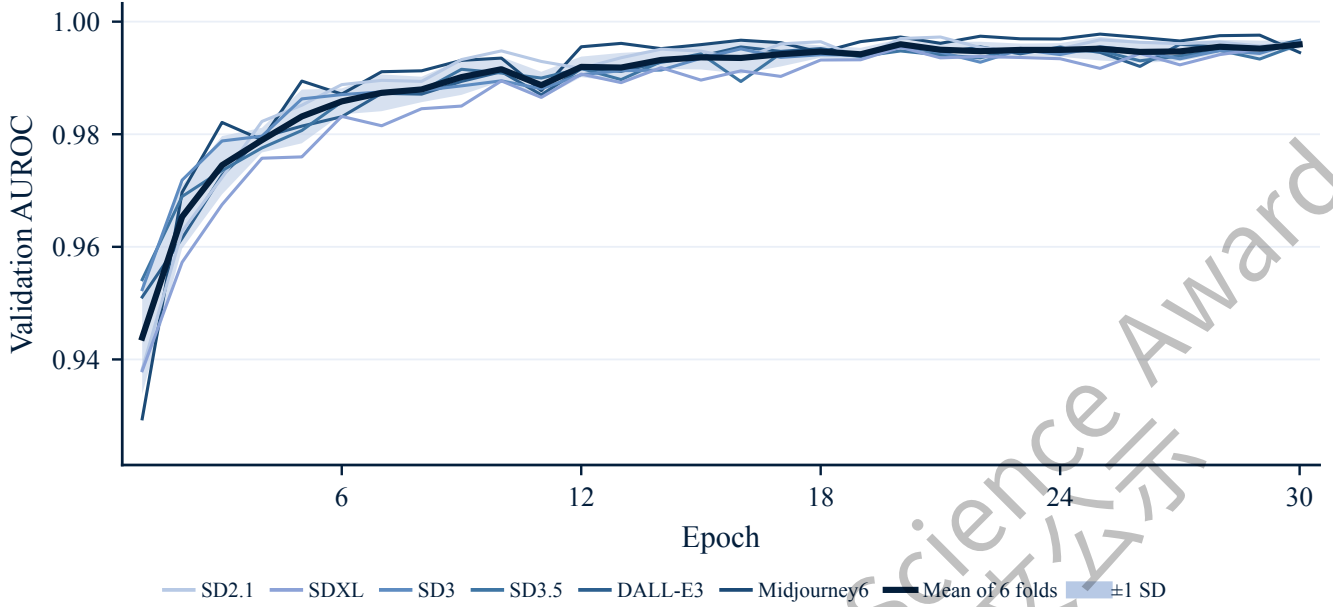


Figure 8. Validation AUROC on the visible generators; curves and band as in Fig. 6. The checkpoint of each fold is selected by this curve.

Table 4. Six-fold LOGO results on COCO_AI (%). Generator columns report held-out AUROC; Mean denotes mean AUROC $\pm$ sample SD across sources (denominator $n - 1$, not multi-seed statistics), Min the minimum AUROC, and TPR_1 the mean $\text{TPR}@1\%\text{FPR}$. Params. counts all parameters in millions, including training auxiliaries. Bold marks the lowest parameter count and highest summary scores.

Method	Params.	Mean	Min	TPR_1	MJ6	DALL-E3	SD3.5	SD3	SDXL	SD2.1
CNNSpot-ResNet50 ¹⁰	23.51	90.11 ± 3.21	86.70	30.98	88.33	88.48	86.70	89.82	95.76	91.56
ConvNeXt-T ¹²	27.82	79.38 ± 5.83	74.10	12.55	74.10	88.58	76.26	78.37	84.30	74.63
Swin-T ¹¹	27.52	65.87 ± 9.33	49.85	3.87	49.85	74.31	61.12	65.79	74.07	70.07
CLIP ViT-B/16 ¹³	86.19	75.85 ± 5.14	71.67	11.66	72.80	84.91	74.35	71.67	78.93	72.43
DINOv2 ViT-S/14 ¹⁴	22.06	81.77 ± 6.00	75.46	12.46	80.84	89.41	78.89	75.46	88.90	77.13
NPR ⁵	1.44	90.03 ± 13.25	71.63	55.99	71.63	74.33	98.49	99.02	99.41	97.30
FerretNet ²⁵	1.06	81.88 ± 26.24	47.33	55.54	47.33	48.74	97.56	99.74	99.90	98.02
Ours	1.64	95.63 ± 4.56	88.47	57.63	91.50	88.47	97.68	99.33	99.52	97.31

considerably when transferred to DALL-E3 and Midjourney6, where FerretNet falls to 48.74% and 47.33%. Our method remains above 97% on all four Stable Diffusion sources (97.31%, 99.52%, 99.33%, and 97.68% on SD2.1, SDXL, SD3, and SD3.5) and attains the highest AUROC of 91.50% on Midjourney6, 3.18 points above the next best result, indicating that the joint representation of multi-response forensic signals and RGB content signals can cover part of the transferable structure under source variation, rather than merely amplifying one class of easily detectable generation trace.

Under low-false-positive conditions, the mean TPR_1 of our method is 57.63%, higher than the 55.99% of NPR and the 55.54% of FerretNet, and the source-wise behavior in Fig. 10 follows the same pattern: the AUROC stays within 0.94 percentage points of the strongest baseline on five of the six generators and leads by 3.18 points on Midjourney6, while TPR_1 reaches 91.77% on SDXL, 89.68% on SD3, 70.47% on SD3.5, and 61.35% on SD2.1. This advantage does not extend to every source: the AUROC on DALL-E3 is slightly below that of DINOv2,

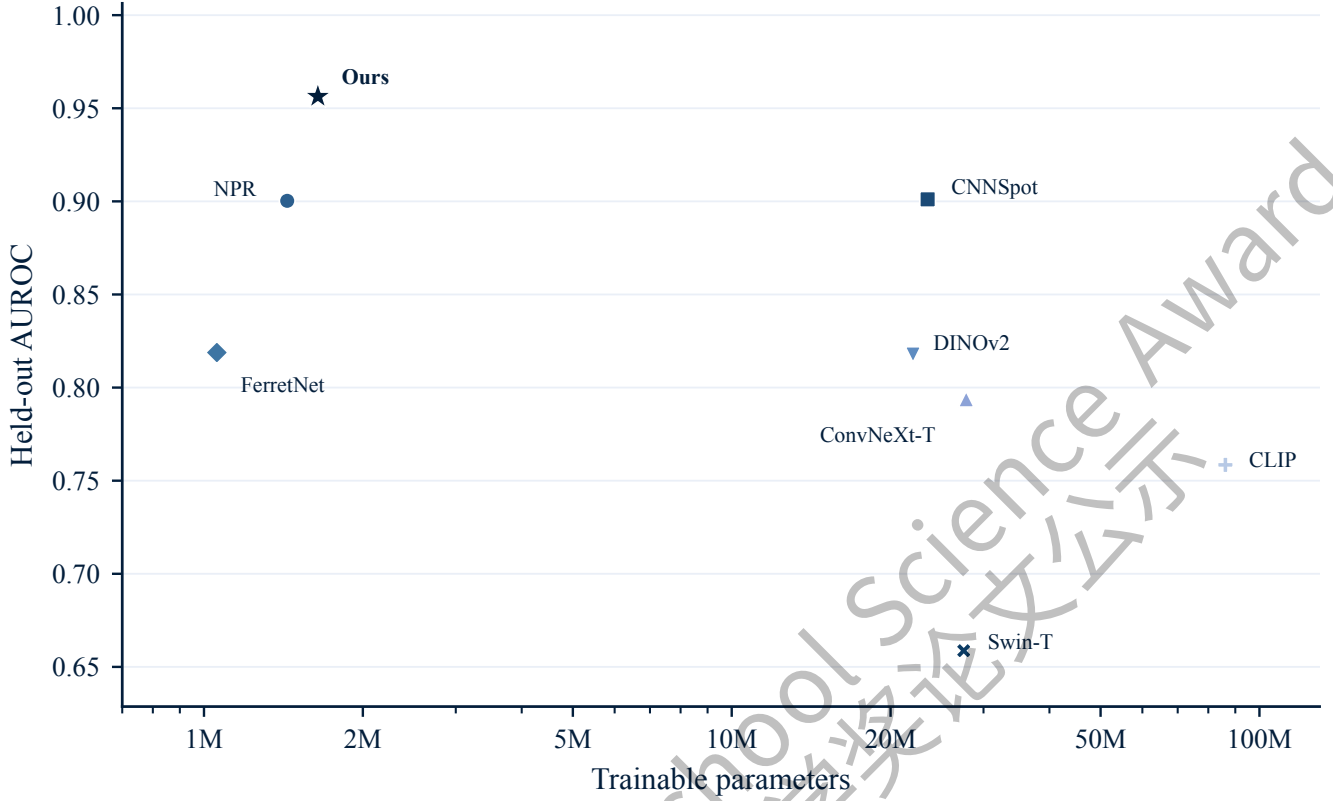


Figure 9. Accuracy–parameter frontier on the six held-out generators. Each marker is one detector of the main comparison and the horizontal axis is logarithmic; our model attains a mean AUROC above 95% at 1.64M parameters, whereas the other detectors below two million parameters remain below 91%.

and TPR_1 drops to 13.27% there and 19.25% on Midjourney6, indicating that the improvement in overall ranking capability has not fully translated into high-confidence detection on those sources. In screening practice, a fixed false-alarm budget determines how many real images may be flagged, and TPR_1 measures how many generated samples are detected under that budget. Overall, the main advantage of our method is that it improves the cross-source average performance and mitigates the degradation on the worst source at a relatively small parameter scale: against CNNSpot the margin reaches 10.98 points on SD3.5 and 9.51 points on SD3, and against the mean of the seven baselines it reaches 22.23 points on Midjourney6.

4.5 Validation–Held-out Gap

Fig. 11 relates the mean validation AUROC on the visible generators to the mean held-out AUROC of the same runs, for the detectors of Table 4. On the visible generators the comparison is saturated: six of the eight detectors reach at least 89% validation AUROC, and the two highest scores belong to FerretNet (99.84%) and our model (99.70%). The held-out generators reorder the field. Our model leads with 95.63%, followed by CNNSpot (90.11%) and NPR (90.03%), whereas FerretNet drops to 81.88% and ranks fourth, although it is first on validation; measured against our model its held-out deficit is 13.75 percentage points, and its validation-to-held-out drop of 17.96 points is the largest of the eight detectors.

A small drop is not by itself evidence of transfer. Swin-T has the smallest one (1.59 points), but it starts from a validation AUROC of 67.46%, and its held-out score of 65.87% is the lowest of the table. Among the four detectors above 95% validation AUROC, our model retains the most: its gap is 4.06 points, against 6.69 for CNNSpot, 9.30 for NPR, and 17.96 for FerretNet, and its validation AUROC is also the steadiest across folds (sample SD 0.09 points, against 0.13 to 2.22 points for the other detectors). Validation on the visible generators thus measures how

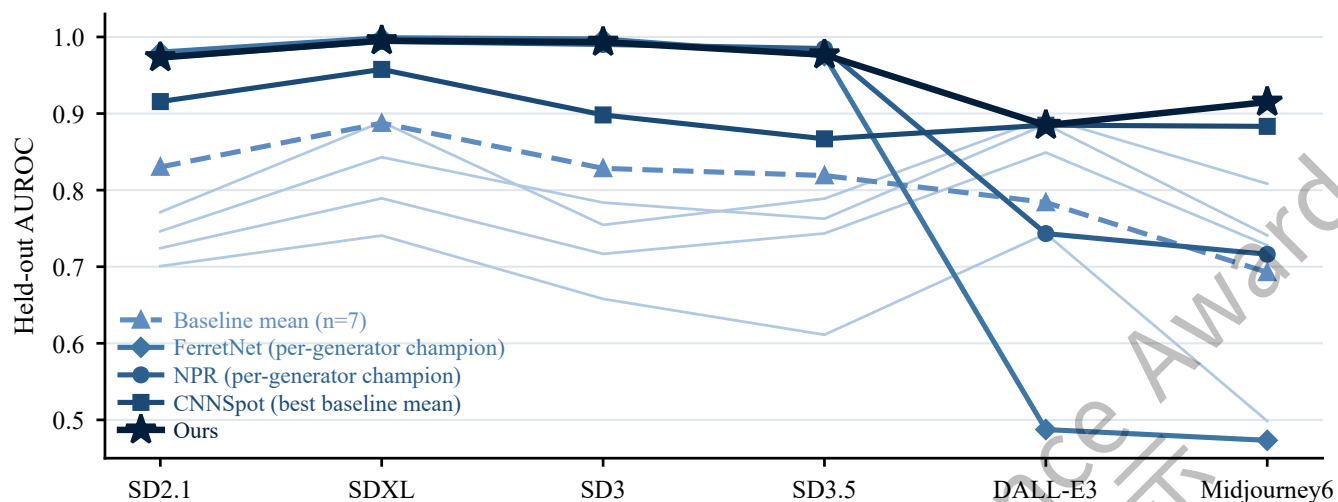


Figure 10. Per-generator advantage over the strongest competitors. Held-out AUROC per generator for our model, for the two lightweight baselines that act as per-generator champions on the Stable Diffusion sources (FerretNet and NPR), for the baseline with the best mean (CNNSpot), and for the baseline mean, with the individual baselines in light lines.

well a detector fits the generation processes it has seen, rather than how much of that fit survives a change of source, and the comparisons above are decided by held-out retention; this is why the results are reported per source beside the mean and the worst source instead of as a single validation-level number.

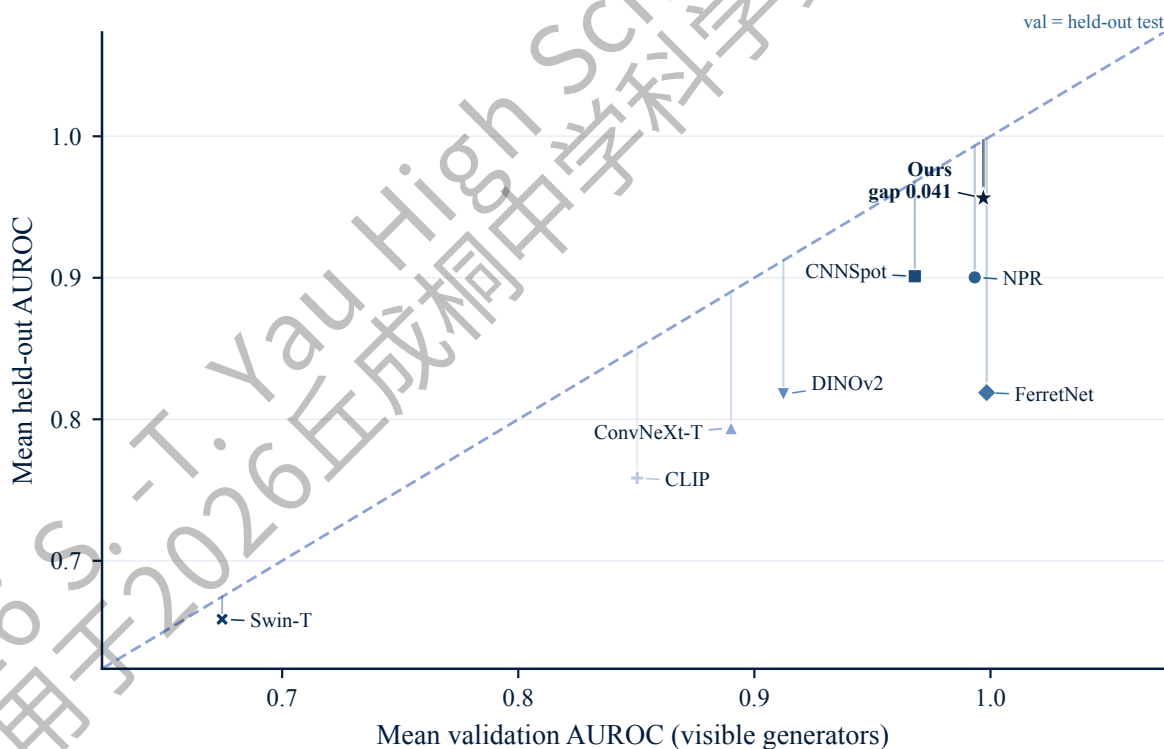


Figure 11. Validation versus held-out AUROC. Each marker is one detector of Table 4: the horizontal axis is the mean validation AUROC on the visible generators, the vertical axis the mean held-out AUROC over the six leave-one-generator-out folds, and the dashed line marks equal values, so the vertical distance to it is the accuracy lost when the source changes. Validation never uses the held-out generator. Error bars are omitted, as the spread across folds is dominated by which generator is held out and is reported per generator in Fig. 10.

4.6 Ablation Studies

Table 5. Six-fold structural ablation (%); metrics follow Table 4. Full denotes the complete model retrained for ablation, distinct from the main-table run.

Variant	Mean	Min	TPR ₁
RGB-only	88.48 $\pm$ 4.43	83.50	22.43
UFEB-only	93.20 $\pm$ 8.74	78.25	53.09
w/o Hybrid2D	94.92 $\pm$ 6.35	83.46	58.48
Full	95.34 $\pm$ 5.19	86.81	57.07

RGB-UFEB complementarity. The full model improves mean AUROC over RGB-only and UFEB-only by 6.86 and 2.14 percentage points, respectively (Table 5). The two single-branch settings exhibit different source preferences: RGB-only reaches 95.57% on DALL-E3, where UFEB-only attains 78.25%, whereas UFEB-only reaches 97.80% on SD2.1, above the 86.35% of RGB-only. Although the full model does not exceed the single-branch variants on every source—on DALL-E3, RGB-only reaches 95.57% against 86.81% for the retrained full model—its minimum AUROC exceeds that of both single-branch settings. This result supports interpreting the complementarity as a source-preference difference between the two signal representations: RGB preserves content and overall texture information, whereas UFEB provides local scale, neighborhood, and directional responses, so that joint training can reduce the source dependence that arises when a single signal dominates. Fig. 12 shows the per-generator structure behind this aggregate result: the leading variant changes from source to source, and the complete model is the only variant whose worst-generator AUROC stays above those of both single-branch settings.

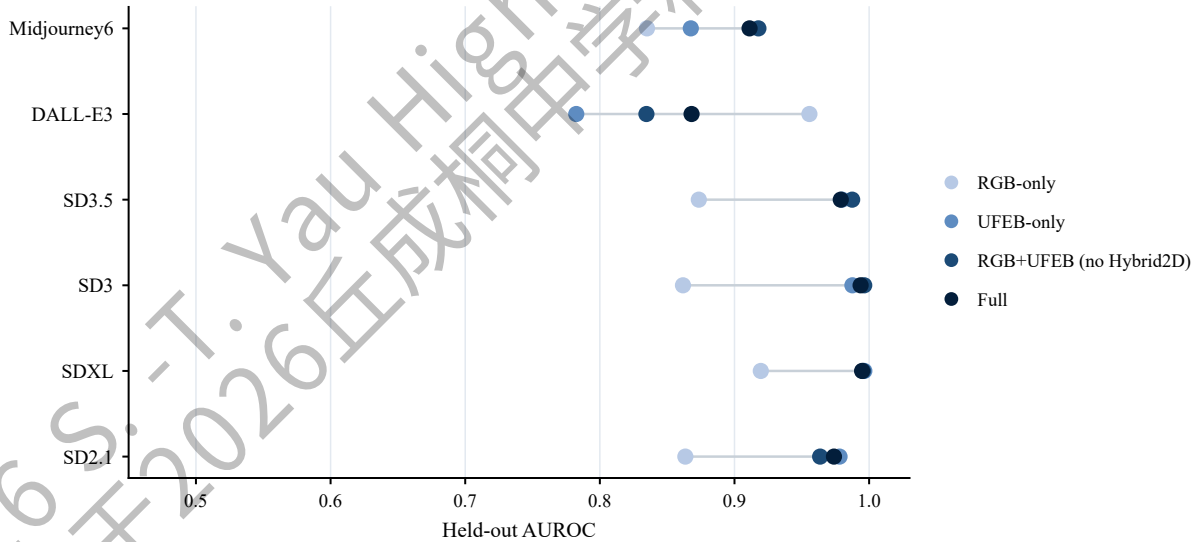


Figure 12. Per-generator view of the structural ablation. Each row is one held-out generator and each marker one variant, joined by a light segment that spans the range of the four variants on that generator.

Hybrid2D high-level integration. Removing high-level integration reduces the mean AUROC from 95.34% to 94.92% and the minimum AUROC from 86.81% to 83.46% (Table 5). This change indicates that continuing to integrate local structure and cross-region information after dual-evidence fusion mainly improves discriminative performance on the weaker sources. However, the model without high-level integration has a higher mean TPR₁, showing that the benefit of this module lies primarily in global ranking and worst-source performance rather than in

a uniform improvement across all evaluation metrics. Fig. 13 collects the four variants in one view: the full model attains the best mean and worst-generator AUROC, while the variant without integration keeps the highest mean TPR@1%FPR.

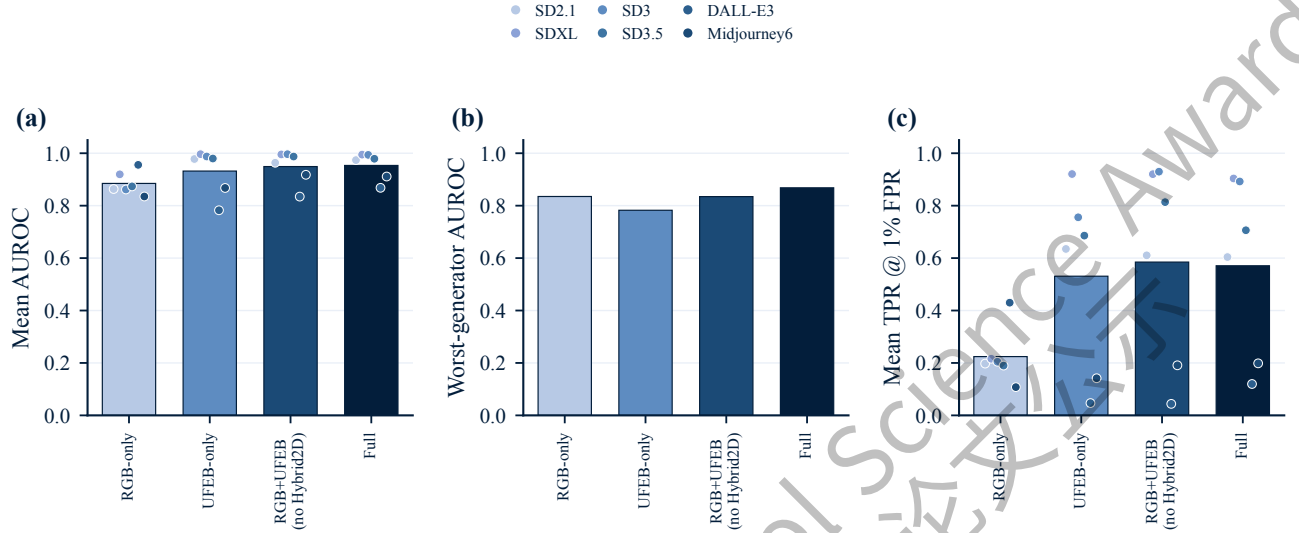


Figure 13. Structural ablation summarized per generator. (a) Mean AUROC, (b) worst-generator AUROC, and (c) mean TPR@1%FPR for the four variants, with the six held-out generators drawn as individual points; the values reproduce Table 5.

Table 6. UFEB response ablation AUROC (%). Mean reports mean $\pm$ sample SD across the three sources. Response symbols follow Eq. 1; Full uses the corresponding folds from Table 5.

Variant	DALL-E3	MJ6	SD3.5	Mean
w/o $\mathbf{R}^s$	76.81	93.72	98.31	89.62 ± 11.32
w/o $\mathbf{R}^c$	76.83	92.52	98.53	89.29 ± 11.21
w/o $\mathbf{R}^d$	87.48	87.91	96.24	90.54 ± 4.94
Full	86.81	91.12	97.89	91.94 ± 5.59

UFEB response composition. Table 6 examines the contribution of the three classes of local cue by zeroing a specified response while keeping the encoding structure unchanged. The full UFEB attains a mean AUROC of 91.94% across the three sources; removing the scale, contextual, or directional response lowers this by 2.33, 2.65, and 1.40 percentage points, respectively. Removing the scale and contextual responses mainly affects DALL-E3, whereas removing the directional response degrades performance on Midjourney6 and SD3.5. Some simplified combinations are better on individual sources, but the full combination attains the highest mean AUROC, showing that the value of the three responses lies in cross-source complementarity rather than in providing an identical gain on every source.

Fig. 14 illustrates the complementarity of the two classes of signal at the sample level. In the first row, the RGB branch fails to yield a generated-image decision, whereas the UFEB branch produces a high-confidence response to local structural anomalies; the second row shows the opposite case, in which the RGB branch can reach a decision

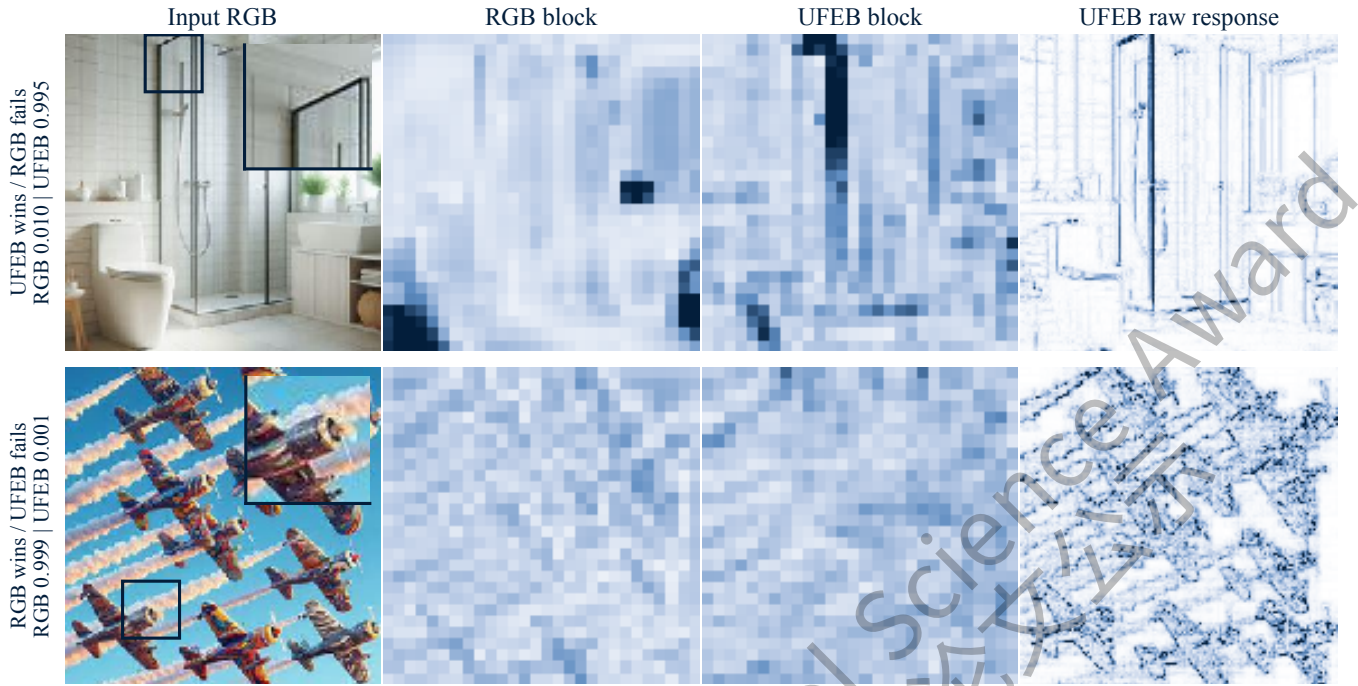


Figure 14. DALL-E3 examples selected for maximal opposite RGB-UFEB probability differences. Scores are generated-image probabilities from the same frozen head with the other branch removed. Black boxes locate peak UFEB response RMS; map scales differ across columns.

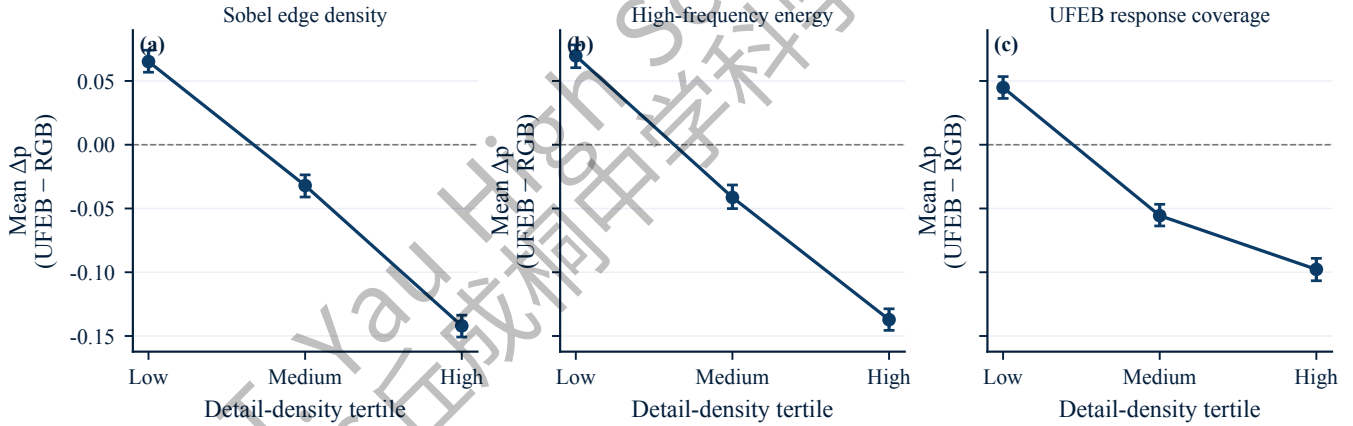


Figure 15. Mean branch-intervention probability difference $\Delta p = p_{\text{UFEB}} - p_{\text{RGB}}$ across detail-density tertiles. Panels use Sobel edge density, high-frequency energy, and UFEB response coverage, respectively; error bars indicate bootstrap 95% confidence intervals.

from overall appearance and spatial layout, while the UFEB responses are comparatively dispersed and insufficient for effective discrimination on their own. Local forensic responses are therefore not a simple duplicate of the RGB signal, and preserving the global RGB observation path is necessary to avoid insufficient selectivity of the local responses. A frozen-model intervention study with an amplitude-spectrum control points the same way (Fig. 16): with the trained weights frozen, keeping only the RGB pathway yields a mean AUROC of 73.61% and keeping only the UFEB pathway 85.59%, whereas a fixed FFT-amplitude representation reaches only 30.32% and collapses to 3.74% on Midjourney6; with both pathways active the same protocol attains 91.73%, the best of the four settings.

To examine whether the above observation is confined to individual samples, we further stratify by detail density over the 24,060 test images of the six held-out generators (Fig. 15). All three density definitions show Δp changing

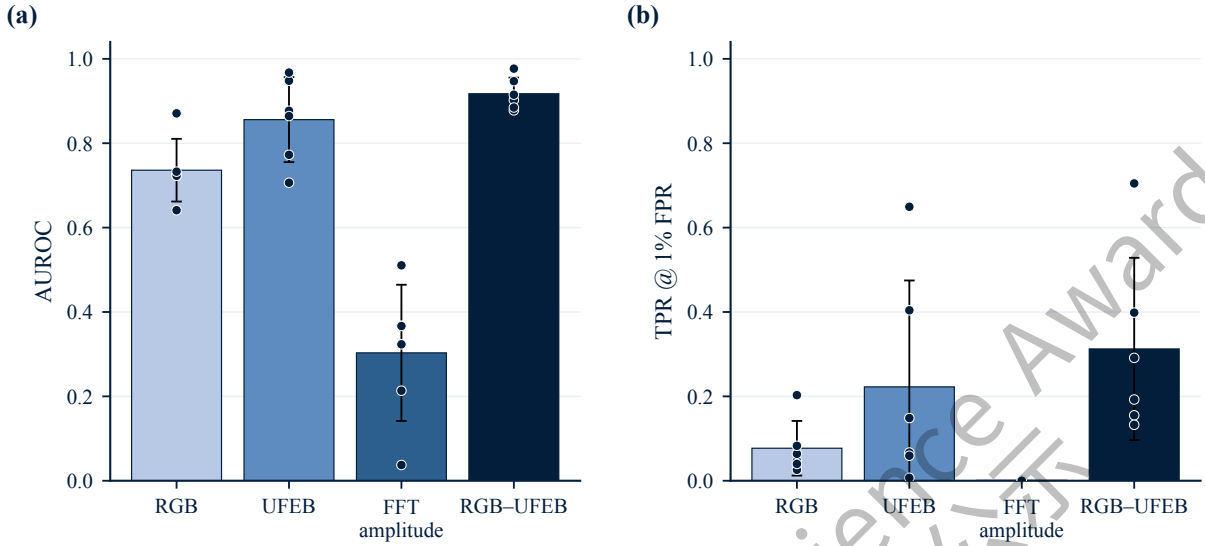


Figure 16. Frozen-model branch intervention with an amplitude-spectrum control. With the trained weights frozen, the RGB pathway alone, the UFEB pathway alone, a fixed FFT-amplitude representation, and the intact model are compared on the six held-out generators; (a) AUROC and (b) TPR@1%FPR. Only pathway interventions are performed and no retraining is involved, so these numbers are not directly comparable with the retrained variants of Table 5.

from positive values in the low-detail group to negative values in the high-detail group, falling from $+0.065$, $+0.070$, and $+0.045$ to -0.142 , -0.137 , and -0.098 , respectively. This indicates that when local detail is sparse, UFEB contributes more through relatively concentrated forensic responses, whereas the relative contribution of RGB appearance and overall structure rises as detail becomes spatially denser. Under Sobel grouping, the AUROC of the UFEB pathway still falls from 0.853 in the low-detail group to 0.745 in the high-detail group, while the RGB pathway remains around 0.70 across the three groups; the experiment therefore supports a change in the relative contribution of the two classes of evidence with signal organization, rather than RGB already surpassing UFEB in overall performance under high-detail conditions. The sample dependence of the two classes of evidence provides empirical motivation for gated residual fusion.

4.7 Sensitivity to the Gate Bias

The gate of Sec. 3.3 is initialized with a bias of -2 , an injection strength of $\sigma(-2) \approx 0.12$, which starts training close to an RGB-only solution. To check that this value is not critical, we retrain the model with four biases (-3 , -2 , -1 , and 0 , that is, initial injection strengths from $\sigma(-3) \approx 0.05$ to $\sigma(0) = 0.5$) on three held-out generators: SD3.5, Midjourney6, and DALL-E3. Each point of Fig. 17 is a separate training run with a single seed and with the same protocol and settings apart from the bias, so the values characterize the sensitivity sweep itself rather than reproducing the folds of Table 4.

Fig. 17(a) shows that validation is flat across the sweep: the dashed curves stay between 99.5% and 99.8% at every setting, so the differences caused by the gate bias appear entirely on the held-out side, and the drop in panel (b) mirrors the held-out curves of panel (a). Across the range -2 to 0 the mean AUROC over the three generators varies by less than one percentage point (89.79 to 90.56%) with no monotonic ordering, and the SD3.5 fold is essentially invariant (97.74 to 97.95%), so these three settings are practically equivalent. Only the strongest suppression, -3 , degrades the model: the mean AUROC falls to 87.46% , driven by DALL-E3 dropping to 76.50% against 82.25 to 85.00% for the other settings, whereas Midjourney6 (88.29%) and SD3.5 (97.60%) remain in their usual range; the mean drop over the three generators is 12.15 points at -3 against 9.1 to 9.9 points elsewhere. Since every entry is a single run, differences of about one percentage point lie within run-to-run variation; what stands out is that

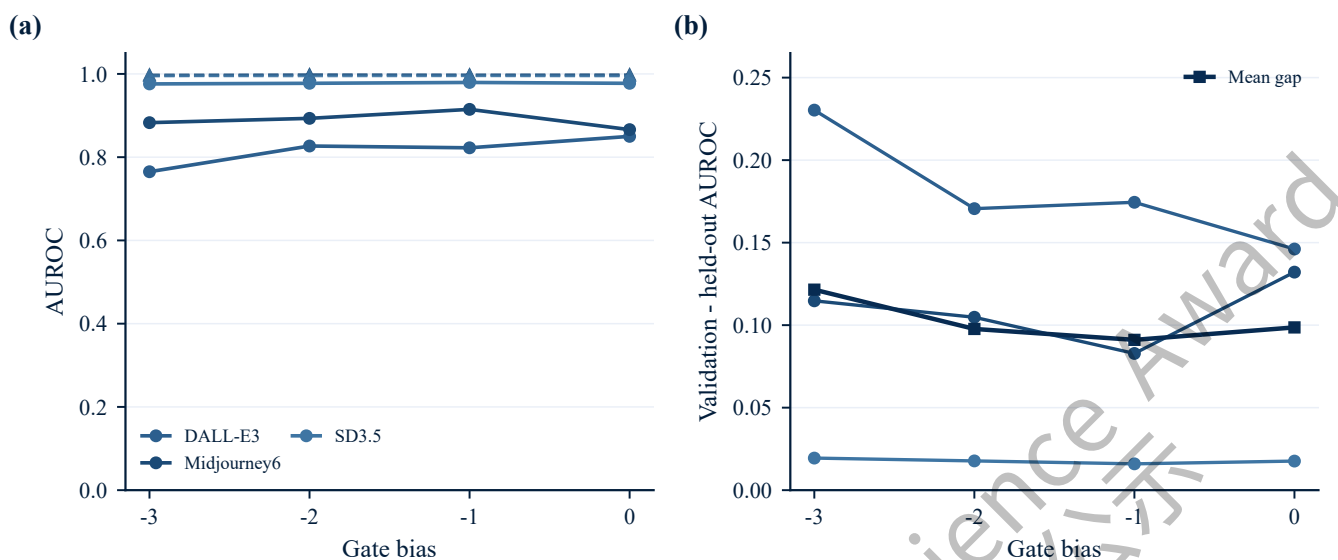


Figure 17. Gate-bias sensitivity. (a) Validation (dashed) and held-out (solid) AUROC for the three generators at each initial gate bias; (b) the resulting drop, with the heavy line marking the mean over the three generators. Every point is a separate training run with a single seed under the same protocol, and only the bias differs; the default bias of the model is -2 , and the held-out generator is never seen during training.

over-suppressing the forensic pathway at initialization costs accuracy, while the exact bias is not critical, and the default of -2 lies inside the stable region.

5 Discussion

The results support a view of cross-generator detection in which the decisive property of a detector is not the capacity of its content representation but the way generation evidence is measured and introduced into the decision. Our model reaches its accuracy with 1.64M parameters, and almost four fifths of that budget is spent on integrating the fused representation, while the construction of the forensic evidence itself adds no parameters. The frozen interventions of Sec. 4.6 make the division of labour visible: measured with the same head, the local evidence alone and the content path alone retain only part of the accuracy that the two reach together, and a fixed amplitude-spectrum control performs well below both. Two properties explain this behaviour: the evidence is measured by operators that do not depend on training, so it cannot drift toward the training generators, and the decision path consults that evidence per image through a learned gate, so the content representation is never forced to carry the forensic signal.

The second observation is that the useful evidence is source- and image-dependent rather than universal. The two branches have different source preferences, the three response families matter in different proportions on different sources, and the relative contribution of the two classes of signal changes systematically with the spatial organization of detail; the gate-bias sweep showed in addition that the outcome is insensitive to a wide range of initial injection strengths, so the model is not balanced on a narrow operating point. Robustness across generators therefore appears to rest on arbitration between several imperfect views rather than on committing in advance to a single artifact family, and per-source and worst-source results are where the differences between detectors actually appear.

Practically, the method combines a small parameter count with a lean inference path. With 1.64M parameters it is more than an order of magnitude smaller than the strongest content-based baseline of the main comparison; inference consumes a single RGB image with no prompts, group structure, or generator labels, because the private branch and the two generator classifiers are discarded after training; and the forensic evidence is computed by fixed operators whose cost is linear in the number of pixels and which remain inspectable independently of the learned encoder. The audited dataset properties support the interpretation of the numbers: after the shared preprocessing

chain, neither the JPEG encoding nor the native image dimensions separate the two classes, so the reported accuracy reflects local generation traces rather than storage or geometry shortcuts.

Several limitations qualify these conclusions. The evaluation covers six diffusion-based text-to-image generators, and transfer to GAN-based synthesis, to editing and inpainting pipelines, to autoregressive image models, and to video forensics is untested. Because the operators of UFEB are fixed, the coverage of the response bank matters: systems whose pipeline leaves different local signatures fall outside what the present study can claim, and a generator-level sensitivity analysis that measures each response family on images from new pipelines would make this question directly testable. The preprocessing chain removes the resolution cue deliberately, but deployment-time recompression and resizing chains are therefore represented only in part by the JPEG and blur perturbations used during training.

On the two hardest sources, DALL-E3 and Midjourney6, the ranking quality of the model is not matched by its behaviour at low false-alarm rates, and the validation-held-out analysis explains why this is difficult to repair by selection alone: validation on the visible generators saturates for all competitive detectors and carries little information once the model is well trained. Two effects are worth separating here: the model’s held-out retention is the best of the compared detectors, so the remaining weakness is a limit of the absolute detection level on those sources rather than a collapse under source shift; and separate runs with nominally identical settings, in the gate-bias sweep, differed by nearly three points on DALL-E3 while landing at the same validation level. Performance at the hard end therefore depends on training variation that the current protocol can measure only through held-out results, and multi-seed evaluation with variance-aware selection is needed before small differences on those sources can be treated as reliable.

The evaluation practice behind these results carries a methodological implication as well: because validation saturates, ranking detectors by validation scores or by a single cross-generator average can invert the ranking that held-out retention produces, and the audits that exclude storage- and geometry-based shortcuts are a precondition for attributing accuracy to generation traces rather than to the benchmark construction.

Methodologically, four choices limit the present design and indicate its natural extensions. First, UFEB is deliberately parameter-free, which makes the evidence stable and auditable but leaves the response bank fixed; learnable residual filters constrained by fixed-operator priors, or a larger basis from which responses are selected during training, are natural ways to extend coverage beyond the three families measured here, and the response ablation indicates the value at stake. Second, evidence is injected at a single stage through a single gate; multi-stage injection and resolution-specific gates are unexplored, although the useful scale of evidence varies with image organization. Third, the model is a binary detector that neither localizes manipulated regions nor attributes the generator; the private representation separates the visible generators during training, but whether it supports attribution of unseen ones remains open. Fourth, the five terms of the objective were not ablated individually and their weights were fixed by a small pilot search, so which part of the objective actually carries the transfer remains to be isolated.

Robustness to deliberate evasion is outside the scope of this study: adversarial perturbations, laundering through a generator or an encoder, and strong recompression were not evaluated, and evidence computed by fixed operators is not obviously invariant to them. The broader implication is that cross-generator detection benefits from separating the measurement of generation evidence from the decision about that evidence: measurements that do not learn are a stable source of local signal, and the learned part of the model is best spent deciding where that signal matters. The remaining gaps, namely harder and more diverse generators, deliberate evasion, and the localization and attribution tasks that a binary verdict does not answer, define the agenda that follows.

6 Conclusion

AI-generated images are now produced in everyday quantities, and their detection has become a routine requirement for media forensics and information security rather than a laboratory exercise. As new text-to-image systems appear faster than detectors can be retrained on them, the practically relevant question is not how well a detector fits

the generators it has seen, but how much of its behaviour survives a change of source, especially under the strict false-alarm budgets of security screening. The difficulty is that generation traces are properties of the synthesis pipeline rather than of the image content, so a detector that binds to the artifacts of its training sources loses accuracy exactly where it is needed most. This work approached the problem by treating the RGB observation and the local traces of the generation process as two complementary sources of evidence, and by leaving it to the decision layer, rather than to the design, to decide how much of each should enter a verdict.

The proposed Unified Forensic Evidence Block measures scale consistency, local contextual deviation, and directional detail with fixed signed operators, adding no trainable parameters and no dependence on the training generators. Its responses are encoded separately from the RGB path and injected into it through a gated residual whose strength starts close to zero, so the content representation stays intact while local evidence is added per location and channel; the gate therefore has to learn to rely on the forensic signal, and the parameters of the model concentrate in the integration stages rather than in the evidence measurement. Hybrid2D then integrates the fused representation over local structure and row- and column-wise relations, and the training objective organizes the shared and private representations so that source-related information is kept away from the representation that makes the decision. Prompt groups and generator labels are used only during training; at inference the model consumes a single RGB image and needs no metadata.

Under six-fold leave-one-generator-out evaluation on COCO_AI, the model attains a mean AUROC of 95.63%, a worst-source AUROC of 88.47%, and a mean TPR@1%FPR of 57.63%, exceeding the strongest competing detector by 5.53 percentage points on average while using more than an order of magnitude fewer parameters. It stays above 97% AUROC on all four Stable Diffusion sources, leads on Midjourney6 by 3.18 points, and keeps the smallest validation-to-held-out drop among the strongest detectors (4.06 points, against 6.69 to 17.96 points for the closest competitors), which shows that the advantage is retained ranking ability rather than a single easy case. The ablations attribute the behaviour to the two evidence streams and their combination: measured with the same frozen head, the content path alone and the local evidence alone retain 73.61% and 85.59% mean AUROC, the two together reach 91.73%, and a fixed frequency-amplitude control collapses to 30.32%, while removing any single response family costs between 1.40 and 2.65 points. The detail-density analysis completes the picture of when each stream matters, showing the local evidence contributing most where detail is sparse and relatively less as detail density rises; the training and gate-bias analyses indicate a stable optimization with no overfitting signature in the observed range and no sensitivity to the initial injection strength within a broad interval; and the dataset audits and the validation-held-out comparison confirm that the reported numbers are not decided by storage format, native resolution, or validation fit.

The remaining weaknesses define the next steps. At low false-alarm rates the detector is still limited on the hardest sources, DALL-E3 and Midjourney6, and the present study covers six diffusion-based generators with a single training seed per configuration; extending the benchmark to GAN-based synthesis, editing pipelines, autoregressive models, video, and realistic post-processing chains, together with multi-seed and variance-aware selection, is necessary before stronger deployment claims can be made. On the method side, the fixed response bank can be complemented by learnable, constrained operators, the injection can be extended to several stages, and the private representation can be directed toward generator attribution or localization, which the present binary verdict does not provide; robustness against deliberate anti-forensic manipulation also awaits systematic evaluation. Given the residual error rates on the hardest sources, deployment is best treated as one piece of evidence inside a broader forensic workflow rather than as a standalone verdict.

Taken together, the study points to two practical directions for cross-generator forensics: evidence that is measured without learning provides a stable basis for decisions under source shift, and evaluating detectors by retained per-source behaviour, rather than by validation fit or a single average, is what makes such claims testable. The lightweight design and the shortcut audits of this work are one instance of that combination, and they offer a starting point for extending cross-generator detection toward more diverse generators and stricter deployment conditions.

References

1. Sha, Z., Li, Z., Yu, N. & Zhang, Y. DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 3418–3432, DOI: 10.1145/3576915.3616588 (2023).
2. Verdoliva, L. Media forensics and DeepFakes: An overview. *IEEE J. Sel. Top. Signal Process.* **14**, 910–932, DOI: 10.1109/JSTSP.2020.3002101 (2020).
3. Durall, R., Keuper, M. & Keuper, J. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7887–7896, DOI: 10.1109/CVPR42600.2020.00791 (2020).
4. Frank, J. *et al.* Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, 3247–3258 (2020).
5. Tan, C. *et al.* 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*, 28130–28139, DOI: 10.1109/CVPR52733.2024.02657 (2024).
6. Wang, Z. *et al.* DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22388–22398, DOI: 10.1109/ICCV51070.2023.02051 (2023).
7. Ricker, J., Lukovnikov, D. & Fischer, A. AEROBLADE: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9130–9140, DOI: 10.1109/CVPR52733.2024.00872 (2024).
8. Cozzolino, D., Poggi, G., Nießner, M. & Verdoliva, L. Zero-shot detection of AI-generated images. In *European Conference on Computer Vision*, 54–72, DOI: 10.1007/978-3-031-72649-1_4 (2024).
9. Zhu, M. *et al.* GenImage: A million-scale benchmark for detecting AI-generated image. In *Advances in Neural Information Processing Systems*, vol. 36, 77771–77782, DOI: 10.52202/075280-3398 (2023).
10. Wang, S.-Y., Wang, O., Zhang, R., Owens, A. & Efros, A. A. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8692–8701, DOI: 10.1109/CVPR42600.2020.00872 (2020).
11. Liu, Z. *et al.* Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9992–10002, DOI: 10.1109/ICCV48922.2021.00986 (2021).
12. Liu, Z. *et al.* A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11966–11976, DOI: 10.1109/CVPR52688.2022.01167 (2022).
13. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, 8748–8763 (2021).
14. Quab, M. *et al.* DINOv2: Learning robust visual features without supervision. *Transactions on Mach. Learn. Res.* (2024).
15. Ojha, U., Li, Y. & Lee, Y. J. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24480–24489, DOI: 10.1109/CVPR52729.2023.02345 (2023).
16. Liu, H. *et al.* Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10770–10780, DOI: 10.1109/CVPR52733.2024.01024 (2024).
17. Tan, C. *et al.* C2P-CLIP: Injecting category common prompt in CLIP to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 7184–7192, DOI: 10.1609/aaai.v39i7.32772 (2025).
18. Zheng, C. *et al.* Breaking semantic artifacts for generalized AI-generated image detection. In *Advances in Neural Information Processing Systems*, vol. 37, 59570–59596, DOI: 10.52202/079017-1903 (2024).
19. Dong, S. *et al.* Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3994–4004, DOI: 10.1109/CVPR52729.2023.00389 (2023).
20. Fridrich, J. & Kodovsky, J. Rich models for steganalysis of digital images. *IEEE Transactions on Inf. Forensics Secur.* **7**, 868–882, DOI: 10.1109/TIFS.2012.2190402 (2012).
21. Luo, Y., Zhang, Y., Yan, J. & Liu, W. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16312–16321, DOI: 10.1109/CVPR46437.2021.01605 (2021).
22. Liu, H. *et al.* Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 772–781, DOI: 10.1109/CVPR46437.2021.00083 (2021).
23. Chen, S. *et al.* Local relation learning for face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 1081–1088, DOI: 10.1609/aaai.v35i2.16193 (2021).
24. Tan, C., Zhao, Y., Wei, S., Gu, G. & Wei, Y. Learning on gradients: Generalized artifacts representation for

- GAN-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12105–12114, DOI: 10.1109/CVPR52729.2023.01165 (2023).
25. Liang, S., Liu, J., Chen, R. & Guan, Q. FerretNet: Efficient synthetic image detection via local pixel dependencies. In *Advances in Neural Information Processing Systems*, vol. 38, 47079–47103, DOI: 10.52202/085713-1402 (2025).
 26. Jeong, Y., Kim, D., Ro, Y. & Choi, J. FrePGAN: Robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 1060–1068, DOI: 10.1609/aaai.v36i1.19990 (2022).
 27. Tan, C. *et al.* Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 5052–5060, DOI: 10.1609/aaai.v38i5.28310 (2024).
 28. Yang, Z. *et al.* All patches matter, more patches better: Enhance AI-generated image detection via panoptic patch learning. In *International Conference on Learning Representations* (2026).
 29. Tan, H. *et al.* Veritas: Generalizable deepfake detection via pattern-aware reasoning. In *International Conference on Learning Representations* (2026).
 30. Qian, Y., Yin, G., Sheng, L., Chen, Z. & Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, 86–103, DOI: 10.1007/978-3-030-58610-2_6 (2020).
 31. Cheng, S., Lyu, L., Wang, Z., Zhang, X. & Schwag, V. 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*, 13455–13465, DOI: 10.1109/CVPR52734.2025.01256 (2025).
 32. Chen, B., Zeng, J., Yang, J. & Yang, R. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, vol. 235, 7621–7639 (2024).
 33. Yan, Z., Zhang, Y., Fan, Y. & Wu, B. UCF: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22355–22366, DOI: 10.1109/ICCV51070.2023.02048 (2023).
 34. Li, O. *et al.* Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2405–2414, DOI: 10.1145/3690624.3709392 (2025).
 35. Yan, S. *et al.* A sanity check for AI-generated image detection. In *International Conference on Learning Representations* (2025).
 36. Corvi, R. *et al.* On the detection of synthetic images generated by diffusion models. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1–5, DOI: 10.1109/ICASSP49357.2023.10095167 (2023).
 37. Wang, H. *et al.* Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision* (2020).
 38. Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D. & Erhan, D. Domain separation networks. In *Advances in Neural Information Processing Systems*, vol. 29, 343–351 (2016).
 39. Ganin, Y. & Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, 1180–1189 (2015).
 40. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, 1597–1607 (2020).
 41. Khosla, P. *et al.* Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, vol. 33, 18661–18673 (2020).
 42. Guillaro, F. *et al.* 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).
 43. Wang, Q. *et al.* ECA-Net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11534–11542 (2020).
 44. Zhang, R. Making convolutional networks shift-invariant again. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, 7324–7334 (2019).
 45. Ho, J., Kalchbrenner, N., Weissenborn, D. & Salimans, T. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180* DOI: 10.48550/arXiv.1912.12180 (2019).

致谢

本研究围绕生成式人工智能背景下的AI生成图像检测问题展开，重点关注不同图像生成模型之间存在分布差异时，检测模型的跨生成器泛化能力。在研究开展过程中，本人得到了中山市小榄中学“小平科技创新实验室”的支持，并得到方劲老师与文亮老师的悉心指导。在此，谨向两位老师致以诚挚的感谢。

首先，感谢方劲老师在本研究的选题、理论分析和论文写作等方面给予的系统指导。在研究初期，方老师结合当前生成式人工智能与数字图像取证领域的发展，对研究问题的价值、研究边界以及整体技术路线进行了讨论和把关。在论文推进过程中，方老师围绕研究逻辑、实验设计、结果分析以及论文整体结构多次提出指导意见，并重点指导本人如何从实验现象中提炼研究问题、组织论证逻辑，使论文的研究思路和学术表达更加完整、严谨。

同时，感谢文亮老师在论文写作与相关理论方面给予的专业指导。文老师围绕AI生成图像检测、图像特征表示以及模型泛化等相关理论与本人进行了多次讨论，并对论文中的理论表述、方法阐释、实验结果分析和英文论文写作提出了具体建议。这些指导帮助本人进一步理解不同视觉信息和取证特征在AI生成图像检测中的作用，也使研究中的方法设计与理论分析能够形成更加清晰的对应关系。

本研究的模型设计、程序实现、实验运行、数据整理、结果统计、图表制作以及论文撰写均由本人自主完成。研究过程中，本人围绕RGB图像信息与局部取证证据之间的互补关系进行了持续探索，完成相关模型的设计与实现，并通过跨生成器实验检验模型面对未知生成来源时的检测能力。在这一过程中，方劲老师和文亮老师主要承担选题、理论及论文写作方面的学术指导，对研究思路、实验逻辑和论文表达进行审阅和把关，不直接参与具体代码编写、实验运行、数据处理及统计计算。

从最初提出研究问题，到不断调整模型和实验方案，再到分析实验结果并完成论文撰写，本人逐步认识到，AI生成图像检测不仅需要关注模型在单一数据集上的检测性能，还需要进一步考察模型面对未知生成模型时，能否提取相对稳定且具有迁移能力的取证特征。围绕这一问题进行反复实验、分析与验证，也成为本研究逐步形成和完善的重要过程。

本研究严格遵守竞赛关于生成式人工智能工具使用的相关规定。本人在整个研究过程中，仅使用人工智能工具辅助处理LaTeX排版问题，人工智能工具不参与论文任何实质性内容的生成、修改或完善。在使用人工智能工具时，本人均明确限定其使用范围：不得对论文中的任何文字内容、研究观点、实验数据、公式内容、图表信息、研究结论及参考文献信息进行增加、删除、改写或润色，仅允许对LaTeX代码及论文版式进行排版层面的检查和调整。

对于人工智能工具提供的所有排版建议，本人均进行人工复核，并将修改后的文稿与原始内容逐项比对，检查正文、公式、数据、图表、引用及参考文献是否发生变化。只有在确认相关修改完全属于LaTeX语法、编译或版式调整，且未改变论文原有内容和含义后，本人才予以采用。如人工智能工具输出中出现任何涉及文字改写、内容增删、逻辑调整或结果解释的内容，本人均不予采用。论文中的研究内容、实验数据、分析过程、研究结论及文字表达均由本人独立完成并最终确认。

最后，衷心感谢中山市小榄中学“小平科技创新实验室”为本研究提供学习、研究与实验环境，感谢在研究过程中给予本人建议和帮助的老师与同学。特别感谢方劲老师和文亮老师在整个研究过程中给予的耐心指导与无偿帮助。正是在不断讨论、实验、修改和验证的过程中，本研究才得以逐步完善。

参赛学生简历

伍浩俊，中山市小榄中学“小平科技创新实验室”核心成员，长期参与信息技术与人工智能方向的项目学习与科研实践。高一阶段已较为系统地完成信息系统开发与人工智能相关知识学习，具备较扎实的程序设计和工程实践基础，能够独立完成信息系统从需求分析、功能设计到前后端开发与部署的基本流程，并能够围绕实际任务开展神经网络模型搭建、结构调整、参数优化及对比实验。

在人工智能方向的学习中，伍浩俊不仅关注模型的应用效果，也逐步将学习重点从“实现一个能够运行的系统”转向“围绕具体问题开展研究”。他能够根据研究问题查阅和整理相关文献，设计实验方案，对模型结果进行比较与分析，并根据实验现象进一步调整技术路线。目前已具备一定的机器学习、深度学习和计算机视觉研究基础，能够较为独立地完成数据处理、模型训练、实验记录、结果分析与可视化等工作。

在科研写作方面，伍浩俊具有一定的学术研究训练基础，能够使用LaTeX完成论文排版与规范化写作，了解科技论文从研究问题、相关工作、方法设计、实验验证到结果讨论的基本结构，并能够将程序开发和模型实验逐步转化为较为规范的研究成果。

其参与开展的科研项目《“心晴日记”系统：基于OCR-BERT的青少年语言灾难化智能评估研究》，尝试将OCR、自然语言处理与BERT模型应用于青少年文本语言特征分析，通过信息技术解决真实场景中的实际问题。该项目获广东省教育厅主办的2026年粤港澳学生信息科技创新大赛一等奖。

在此基础上，伍浩俊进一步开展人工智能生成内容识别方向的研究，开始关注模型在不同数据来源和未知生成模型条件下的泛化问题，并尝试从模型结构、特征表示和实验验证等角度开展较为完整的研究。这一研究经历使其科研能力逐步由系统开发和模型应用，向问题提出、方法设计、实验验证与论文表达相结合的研究型学习转变。

指导老师简历

方劲，中共党员，教育博士生，计算机科学与技术硕士，全国中小学实验教学能手，教育部全国师生信息素养提升实践活动专家，广东省基础教育教研体系建设高中信息技术学科教研项目（中山）成员，广东省十佳科技优秀辅导员，广东省科技劳动实践优秀指导教师，中山市优秀教师，中山市中小学人工智能教育教师培养计划项目指导团队成员。曾主持（参与）国家级课题2项，省级课题6项，市级课题6项，发表计算机SCI 2篇，国际会议EI论文2篇，发表教育类论文国家级论文3篇，省级论文5篇，拥有2项发明专利，1项软件著作权，编写出版教材《人工智能教育》并通过审定。方劲老师长期致力于高中阶段人工智能教育，指导学生在科技创新领域获各级奖项200余项，其中国家级奖项8项（国一6人次），省级30余人次（省一13人次）。

文亮，中共党员，高中信息技术高级教师，广东省优秀指导教师，中山市优秀教师，中山市优秀指导教师，小榄镇优秀竞赛辅导员。曾主持（参与）省级课题3项，市级课题6项。曾主持（参与）省级课题2项，市级课题2项，发表教育类论文4篇，拥有1项发明专利，1项软件著作权，编写出版教材《人工智能教育》并通过审定。长期致力于高中阶段人工智能教育，指导学生在科技创新领域获各级奖项150余项，其中国家级奖项2项，省级25人次（省一7人次）。