COMPOSITIONAL FINGERPRINTS FOR OPEN-SET AI-GENERATED IMAGE ATTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image attribution is increasingly important for digital forensics, copyright protection, and content accountability. In practice, however, attribution is inherently open-set: generative models evolve rapidly, and test images may come from generators unseen during training. This is particularly challenging because new generators often inherit architectures, reuse generation mechanisms, or share model lineages with known sources, causing their forensic traces to partially overlap. Existing methods typically represent each generator with a holistic fingerprint, making it difficult to distinguish shared traces from source specific organization. We propose compositional source fingerprints, based on the insight that generator identity can be characterized by how reusable latent generative patterns are combined rather than by a single indivisible fingerprint. We instantiate this idea with a Restricted Dynamic Hypergraph (RdH), in which fingerprint atoms represent reusable latent patterns and induce hyperedges among images. Restricted activation constructs compact, image-specific atom combinations, while degree-aware propagation suppresses broadly shared atoms with limited source selectivity. The resulting compositional representations support known-source attribution, unknown rejection, and novel source structure discovery. On five Cross-Era protocols, RdH improves open-set attribution AUROC over the strongest evaluated baselines by 9.89 percentage points, while also substantially improving clustering quality among unseen generators.

1 INTRODUCTION

AI-generated images are increasingly prevalent in online content, creative production, and digital media Durall et al. (2020). Beyond detecting whether an image is synthetic, identifying which generator produced it is important for digital forensics, copyright protection, and content accountability Ding et al. (2021); Song et al. (2024); Xu et al. (2025a). In realistic deployment, however, attribution systems cannot assume that every test image comes from a known source Dhariwal & Nichol (2021); Rombach et al. (2022); Cazenavette et al. (2024). Generative models evolve rapidly, and newly released or fine-tuned generators continually enter the ecosystem Park & Owens (2025); Fei et al. (2026). Forcing images from these new generators into a fixed set of known sources can therefore produce unsupported and misleading attribution decisions Song & Itti (2025); Yang et al. (2025); Zheng et al. (2025). This makes open-set AI-generated image attribution, which requires both recognizing known sources and rejecting unknown ones Bendale & Boult (2016); Girish et al. (2021); Fang et al. (2023), essential for practical provenance analysis.

Open-set attribution is particularly challenging because unknown sources may correspond to generators related to those in the known-source set. New models may inherit architectures, reuse generation mechanisms, or evolve from common open-source lineages, causing their synthesis traces to partially overlap with those of existing generators Liu et al. (2024a); Yang et al. (2025). As a result, an unknown source may exhibit substantial evidence associated with a known source while still differing in other aspects. Existing attribution methods typically represent each source using a holistic fingerprint or a global source-discriminative representation Bui et al. (2022); Yang et al. (2023); Cioni et al. (2024). While effective for distinguishing sources within a predefined candidate set, such representations do not explicitly reveal which latent traces are shared across generators or how those traces are combined Wang et al. (2026b). This can make an unknown source appear overly similar to a known source and lead to misattribution.

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

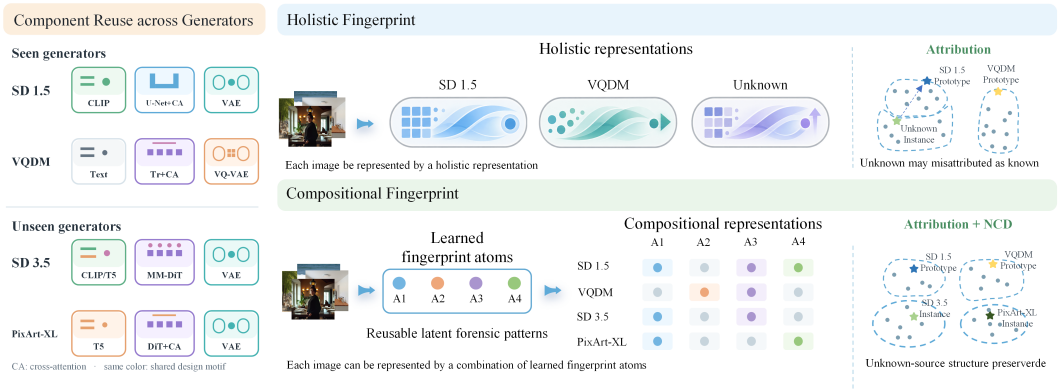


Figure 1: Component reuse across generators can produce shared forensic traces, while holistic fingerprints collapse them into a single representation and obscure how they are organized across sources. Compositional fingerprints instead represent each image through reusable fingerprint atoms and their combinations.

We therefore revisit source fingerprints from a compositional perspective. Our key observation is that a new generator need not introduce an entirely new fingerprint: it may reuse latent generative patterns already present in known sources, but combine them differently. We represent these reusable patterns as generative fingerprint atoms and characterize each image by its atom combination. Individual atoms need not correspond to explicit architectural modules or uniquely identify a source; instead, source identity is reflected by the joint activation pattern of multiple atoms. Thus, sharing some atoms does not imply sharing the same source, while differences in their composition can provide evidence for both attribution and unknown rejection.

Based on this idea, we propose a *Restricted Dynamic Hypergraph* (RdH) for learning compositional source fingerprints. RdH models images as nodes and fingerprint atoms as hyperedges, naturally capturing the many-to-many relations between images and reusable generative patterns. Restricted activation constructs compact, image-specific atom combinations, while degree-aware propagation reduces the influence of broadly shared atoms that provide limited source selectivity. The resulting incidence patterns serve as compositional image features, from which source prototypes are constructed for known-source attribution and unknown rejection. We further examine whether these representations preserve source structure among unknown sources through Novel Class Discovery (NCD) Han et al. (2019); Fini et al. (2021).

We evaluate RdH on five Cross-Era protocols designed to capture diverse shifts in generator evolution. RdH achieves 92.84% average AUROC across the five protocols, outperforming the strongest evaluated baseline by 9.89 percentage points. It also substantially improves clustering quality among unknown sources, supporting the hypothesis that compositional fingerprint structure provides useful evidence beyond holistic source representations.

The main contributions of this paper are summarized as follows:

- We introduce compositional source fingerprints, representing generator identity through combinations of reusable generative fingerprint atoms rather than a single holistic pattern.
- We develop RdH, which combines dynamic restricted image–atom incidence with degree-aware hypergraph propagation to learn compact compositional representations for known attribution and unknown rejection.
- We construct five Cross-Era protocols and demonstrate consistent improvements over strong open-set attribution baselines, including gains of 9.89 percentage points across Cross-Era protocols.

2 Related Work

2.1 Open-Set AI-Generated Image Attribution

AI-generated image attribution aims to identify the generative source of an image Yu et al. (2021); Xu et al. (2025b); Gunn et al. (2025). Source attribution can be supported by active provenance

signals, such as watermarks embedded during generation Fernandez et al. (2023); Wen et al. (2023); Yang et al. (2024), or performed passively by exploiting forensic traces naturally retained in generated images; here, we focus on the latter. Early studies mainly considered a closed-set setting, assuming that training and testing share the same set of candidate generators Marra et al. (2019); Zhang et al. (2019); Sha et al. (2023). Yu et al. (2019) systematically showed that different GANs leave model-specific fingerprints that enable fine-grained source attribution. As generative models continue to evolve, however, the assumption of a fixed candidate set has become increasingly restrictive, motivating attribution in open environments Wang et al. (2020); Ojha et al. (2023); Cozzolino et al. (2024). Girish et al. (2021) explored the detection, clustering, and iterative discovery of GAN models absent from training, while POSE Yang et al. (2023) formulated open-set model attribution as the joint problem of recognizing known sources and rejecting unknown ones. Subsequent studies have addressed generators outside the training set from different perspectives Zheng et al. (2025); Fei et al. (2026); Wang et al. (2023). ASV Abady et al. (2024) performs architecture-level verification using pairwise comparisons or reference images, Cioni et al. (2024) investigate the transferability of general-purpose pretrained features for source attribution, and FSD Nguyen et al. (2025) models multi-scale forensic microstructures to support zero-shot detection, open-set source attribution, and source clustering.

In parallel, recent studies have investigated how attribution systems can exploit limited information from newly encountered generators Fei et al. (2025); Wang et al. (2026a). OCC-CLIP Liu et al. (2024b) formulates origin attribution as few-shot one-class classification, LIDA Wang et al. (2026b) combines retrieval-based attribution with few-shot adaptation, OmniDFA Wu et al. (2026) identifies generators absent from training using a small support set, and IncreFA studies continual adaptation as new generators arrive while explicitly exploiting hierarchical relations among generative architectures. These approaches are particularly useful when reliable reference images or source information become available. In open-world forensic scenarios, however, images may originate from unidentified, inaccessible, or privately fine-tuned generators for which such references cannot be assumed, leaving open-set attribution necessary for recognizing known sources while rejecting samples that cannot be reliably associated with the existing source set Song & Itti (2025). Beyond conventional source-level fingerprints and discriminative embeddings, recent methods have begun to capture richer forensic structure Song et al. (2024), including multi-scale microstructures Nguyen et al. (2025) and architectural hierarchies Qin et al. (2026). Nevertheless, closely related generators may share architectures, generation mechanisms, training configurations, or model lineages, which may lead to partially overlapping learned forensic evidence rather than fully independent source signatures. How such shared evidence is organized differently across generators at the image level, and how this organization can be exploited for open-set attribution, remains comparatively less explored.

2.2 Hypergraph Learning

Hypergraphs provide a natural way to model many-to-many relations, as each hyperedge can connect multiple nodes that share a common latent pattern Zhou et al. (2006); Jiang et al. (2019); Zhou et al. (2023). HGNN extends neural message passing to hypergraph structures Feng et al. (2019), while UniGNN provides a unified formulation for graph and hypergraph propagation Huang & Yang (2021); Yadati et al. (2019). Subsequent work further learns or adapts node–hyperedge associations from input features rather than relying on fixed structures Gao et al. (2022), including soft hyperedge participation in SoftHGNN and shared relational modeling among visual objects in Hyper-YOLO Lei et al. (2025); Feng et al. (2025). These properties make hypergraphs particularly relevant to source attribution, where images from different generators may exhibit overlapping generative traces and therefore participate in shared latent relations. However, existing hypergraph methods typically construct such relations from general feature similarity or visual semantics, without distinguishing source-selective associations from patterns that are broadly shared across many samples. Moreover, repeated propagation can amplify these common relations and make originally distinct representations increasingly similar, motivating approaches that constrain or modify hypergraph diffusion Chen et al. (2022); Li et al. (2025); Zhou et al. (2026). For open-set AI-generated image attribution, the key challenge is therefore not merely to model shared relations, but to construct compact image-specific associations that capture reusable forensic patterns while controlling the influence of broadly activated, weakly discriminative relations during propagation.

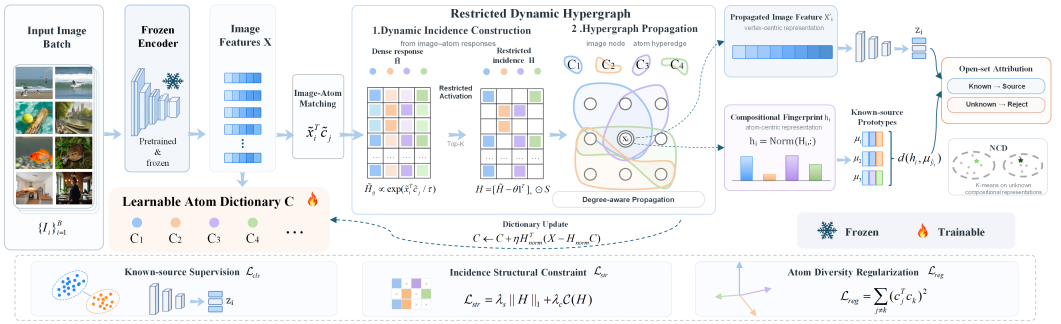


Figure 2: **Overview of RdH.** Image features are matched to a learned atom dictionary through dynamic incidence construction. The restricted incidence matrix supports degree-aware propagation, while its normalized rows provide compositional features for individual images for prototype-based source consistency and auxiliary analysis of structure among unknown sources.

3 METHOD

As illustrated in Fig. 2, RdH represents images as hypergraph nodes and uses a shared set of generative fingerprint atoms to define hyperedges. The weighted connections between images and atoms form the compositional structure. Each row of the incidence matrix records the atom responses of one image, and each column connects images that respond to the same atom. This structure is dynamically constructed from the current inputs and atoms, and is used for feature propagation and compositional representation. We first formulate the open-set attribution problem and then describe how the structure is learned and represented.

3.1 PROBLEM FORMULATION

Let $\mathcal{D}_{\text{tr}} = \{(I_i, y_i)\}_{i=1}^{N}$ denote the training set, where $y_i \in \mathcal{Y}_K$ and $\mathcal{Y}_K$ is the set of sources observed during training. Source labels cover real images and individual generators; the unknown sources in our evaluation are generators not represented in training. At test time, an image may originate from either $\mathcal{Y}_K$ or an unknown source in $\mathcal{Y}_U$, where $\mathcal{Y}_K \cap \mathcal{Y}_U = \varnothing$.

For a test image I_i, the model produces a candidate known source $\hat{y}_i \in \mathcal{Y}_K$ and an unknown score s_i. The image is rejected as unknown when s_i exceeds a selected operating threshold; otherwise, it is attributed to $\hat{y}_i$. Training uses only images and source labels from $\mathcal{Y}_K$, without access to images or labels from unknown sources.

3.2 COMPOSITIONAL STRUCTURE LEARNING

For a mini-batch of B images, a frozen visual encoder produces $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_B]^\top \in \mathbb{R}^{B\times d}$. RdH maintains M generative fingerprint atoms, $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_M]^\top \in \mathbb{R}^{M\times d}$, where atom $\mathbf{c}_j$ defines hyperedge e_j. The dictionary is shared across all known sources, and no source label is assigned to an individual atom.

After ℓ_2 normalization, the preliminary response of image i to atom j is $\widetilde{H}_{ij} = \mathrm{softmax}_j(\bar{\mathbf{x}}_i^\top \bar{\mathbf{c}}_j/\tau)$, where $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{c}}_j$ are the normalized image feature and atom, and τ is a temperature parameter. Row i of $\widetilde{\mathbf{H}}$ describes how image i responds to different atoms, while column j shows how atom j responds across the current images. An image may respond to multiple atoms, and an atom may be shared by multiple images.

We restrict the dense responses in $\widetilde{\mathbf{H}}$ to retain a compact combination for each image. RdH predicts a sample-dependent threshold $\theta_i = f_\theta(\bar{\mathbf{x}}_i)$ and constructs

$$\mathbf{H} = \left[\widetilde{\mathbf{H}} - \boldsymbol{\theta}\mathbf{1}_M^\top\right]_+ \odot \mathbf{S}, \tag{1}$$

where $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_B]^\top$, $[\cdot]_+$ denotes ReLU, and $\mathbf{S} \in \{0,1\}^{B\times M}$ is a row-wise Top-K_H mask computed from the thresholded responses. The threshold removes weak responses while retaining positive responses after threshold subtraction, and $\mathbf{S}$ bounds the number of active connections for

each image. An entry $H_{ij} > 0$ connects image node i to atom hyperedge e_j. Together, the nonzero entries of $\mathbf{H}_{i,:}$ specify the current atom combination of image i. The incidence matrix is recomputed during each forward pass rather than fixed in advance.

The same incidence matrix defines feature propagation among the images in the current batch. We compute the vertex and hyperedge degree matrices as $\mathbf{D}_v = \mathrm{diag}(\mathbf{H}\mathbf{1}_M)$ and $\mathbf{D}_e = \mathrm{diag}(\mathbf{H}^\top \mathbf{1}_B)$. Their diagonal entries are lower-bounded by a small positive constant for numerical stability. The vertex features are updated by

$$\mathbf{X}' = \phi\Big(\mathbf{D}_v^{-\frac{1}{2}}\mathbf{H}\mathbf{D}_e^{-(1+\beta)}\mathbf{H}^\top\mathbf{D}_v^{-\frac{1}{2}}\mathbf{X}\mathbf{\Theta}\Big), \tag{2}$$

where Θ is a learnable feature transformation, $\phi(\cdot)$ is a nonlinear activation, and $\beta > 0$ controls the effect of hyperedge degree. Images that respond to the same atoms exchange information through $HD_e^{-(1+\beta)}H^\top$. For images i and k, atom j contributes in proportion to $H_{ij}H_{kj}/d_{e,j}^{1+\beta}$, so its effect depends on both response strengths. The additional factor $D_e^{-\beta}$ reduces the propagation contribution of atoms activated by many images in the current batch. These atoms remain in the incidence matrix and can still contribute to an image's atom combination. Degree-aware propagation therefore controls how strongly a shared atom mixes image features while retaining its response for compositional representation (see Appendix Sections A.2–A.3).

The current incidence also guides the update of the atoms. Let $\mathbf{H}_{\mathrm{norm}} = \mathrm{RowNorm}(\mathbf{H})$. The batch features are reconstructed as $\mathbf{H}_{\mathrm{norm}}\mathbf{C}$. After the differentiable parameter update for the current mini-batch, the reconstruction residual is assigned to the atoms according to the current incidence weights:

$$\mathbf{C} \leftarrow \mathbf{C} + \eta\,\mathbf{H}_{\mathrm{norm}}^\top\left(\mathbf{X} - \mathbf{H}_{\mathrm{norm}}\mathbf{C}\right), \tag{3}$$

where η is the residual update rate. Each atom is ℓ_2-normalized after the update. The next forward pass recomputes the incidence matrix using the updated atoms, so the atoms and their connections continue to change throughout training. Utility-based dictionary maintenance is also applied during training, with its update rule and parameters given in Appendix B.3. At inference, the atoms and learned model parameters are fixed, while the incidence matrix remains input-dependent.

Known-source labels supervise the propagated features. Let $\mathbf{z}_i = g(\mathbf{X}'_i)$ denote the source classification logits, and let $\mathcal{L}_{\mathrm{cls}}$ be the corresponding cross-entropy loss. For the current mini-batch, we constrain the incidence structure by

$$\mathcal{L}_{\mathrm{str}} = \frac{1}{BM}\left\|\mathbf{H}\right\|_1 - \frac{\alpha}{B}\sum_{j=1}^{M}\frac{1}{d_{e,j}}\left\|\sum_{i=1}^{B}\frac{H_{ij}}{\sqrt{d_{v,i}}}\mathbf{x}_i\right\|_2^2. \tag{4}$$

The first term limits the overall incidence magnitude. The second term aggregates the features connected by each hyperedge, while $d_{e,j}^{-1}$ accounts for differences in its degree. With $a_{ij} = H_{ij}/\sqrt{d_{v,i}}$, the squared aggregation contains the pairwise terms $a_{ij}a_{kj}\mathbf{x}_i^\top\mathbf{x}_k$ between images connected by hyperedge j. The inverse hyperedge degree prevents this quantity from increasing solely because an atom is activated by more images.

Let $\bar{\mathbf{C}}$ denote the row-normalized atom matrix. We use $\mathcal{L}_{\mathrm{reg}} = M^{-2}\|\bar{\mathbf{C}}\bar{\mathbf{C}}^\top - \mathbf{I}_M\|_F^2$ to discourage redundant atom directions. This constraint does not impose semantic independence among the atoms. The complete differentiable objective is $\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{str}}\mathcal{L}_{\mathrm{str}} + \lambda_{\mathrm{reg}}\mathcal{L}_{\mathrm{reg}}$. The residual correction in Eq. equation 3 and utility-based maintenance are update operations during training rather than additional loss terms.

3.3 Compositional Structure Representation

The vertex-centric representation is obtained by propagating features over the current incidence structure. For image i, the images involved in its aggregation and their weights are determined by $\mathbf{H}$, $\mathbf{D}_v$, and $\mathbf{D}_e$. Its atom associations enter the propagated feature $\mathbf{X}'_i$ through these propagation paths and weights. The model predicts a candidate known source using $\mathbf{z}_i = g(\mathbf{X}'_i)$ and $\hat{y}_i = \arg\max_c z_{ic}$.

The hyperedge centric representation describes an image in the coordinates of the generative fingerprint atoms. Since column j of $\mathbf{H}$ corresponds to atom hyperedge e_j, image i is represented by the

normalized incidence row $\mathbf{h}_i = \mathrm{Norm}(\mathbf{H}_{i,:}) = \mathbf{H}_{i,:}/(\|\mathbf{H}_{i,:}\|_2 + \epsilon)$. Its nonzero coordinates identify the atoms activated by the image, and their values give the corresponding response strengths. The incidence structure determines the propagation used to obtain $\mathbf{X}'_i$, whereas $\mathbf{h}_i$ directly records the joint atom activation pattern.

Compositionality here does not require an individual atom to have a predefined semantic meaning. The same atom may appear in images from different sources. Source information is represented by which atoms are activated together and by their relative responses. The source prototypes introduced below summarize recurring joint activation patterns rather than assigning a separate atom to each source. Once the model is fixed, $\mathbf{h}_i$ depends only on image i and the learned atoms; it is unchanged by the other images in the test batch. In contrast, $\mathbf{X}'_i$ depends on the current batch through the hyperedge degrees and feature propagation. After training, we use the fixed model to recompute $\mathbf{h}_i$ for the known-source training images. For each source c, we average these representations and retain the K_p largest coordinates to construct its source prototype:

$$\boldsymbol{\mu}_c = \mathrm{Norm}\left(\mathrm{TopK}_{K_p}\left(\frac{1}{|\mathcal{D}_c|}\sum_{i:y_i=c}\mathbf{h}_i\right)\right), \tag{5}$$

where $\mathcal{D}_c$ contains the training images from source c. TopK_{K_p} retains the original values of the selected coordinates before ℓ_2 normalization. The prototype support size K_p is distinct from the incidence support size K_H used for individual images.

For a test image, $\mathbf{X}'_i$ produces the candidate source $\hat{y}_i$, while the cosine distance $d(\mathbf{h}_i, \boldsymbol{\mu}_{\hat{y}_i})$ measures how closely its atom activation pattern matches the prototype of that source. We combine energy and maximum softmax probability (MSP) uncertainty as $u_{\mathrm{cls}}(\mathbf{z}_i) = \rho\, u_{\mathrm{energy}}(\mathbf{z}_i) + (1-\rho)\, u_{\mathrm{MSP}}(\mathbf{z}_i)$, where $\rho \in [0,1]$, u_{energy} is the sigmoid-scaled energy score, and $u_{\mathrm{MSP}}(\mathbf{z}_i) = 1 - \max_c \mathrm{softmax}(\mathbf{z}_i)_c$. The unknown score is

$$s_i = w_p\, d(\mathbf{h}_i, \boldsymbol{\mu}_{\hat{y}_i}) + (1-w_p)\, u_{\mathrm{cls}}(\mathbf{z}_i), \tag{6}$$

where $w_p \in [0,1]$ balances prototype consistency and classifier uncertainty. This is a reparameterization of the original weights, with $w_e = (1-w_p)\rho$ and $w_m = (1-w_p)(1-\rho)$. At operating threshold γ, the image is rejected as unknown if $s_i > \gamma$; otherwise, it is attributed to $\hat{y}_i$.

Beyond open-set attribution, we use NCD as an auxiliary analysis of the learned representations Zhong et al. (2021). We apply K-means to the compositional representations $\mathbf{h}_i$ of all unknown-source evaluation samples, independently of the rejection decisions. Each evaluation protocol contains four unknown sources, so we set $k = 4$. Unknown-source labels are not used for training, source prototype construction, unknown scoring, or clustering; they are used only after clustering to compute NMI and ARI.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation settings. To evaluate attribution to unknown sources, we construct five Cross-Era settings, denoted S1–S5. Each setting uses real images and generators from the earlier GenImage dataset Zhu et al. (2023) as known sources, while unknown-source images are drawn from the more recent AIGI-Holmes test set Zhou et al. (2025). This temporal separation mimics the continual emergence of new generators. The five settings further vary source relatedness, covering balanced generator compositions, diffusion-dominant known sources, heterogeneous architectures, closely related Janus models, and a U-Net→DiT shift. Thus, S1–S5 represent distinct distribution shifts in source diversity, family relatedness, and generation architecture rather than cross-validation folds. Complete source compositions, sample counts, and supplementary GenImage results are provided in Appendix Sections C and E.1.

Evaluation metrics. We use ACC for known-source attribution, AUROC and FPR95 for unknown-source separation, and open-set accuracy at the median-midpoint threshold (labeled OSCR in the tables) for their joint evaluation. We further apply K-means to unknown-source representations and report NMI and ARI to measure source-structure preservation. The known/unknown test partition is

Table 1: Cross-Era attribution results over five protocols. Each entry reports the macro-average ± standard deviation across protocols. The deviation summarizes cross-protocol variation rather than repeated-run uncertainty. Full protocol-wise results are reported in Appendix E.1.

Method	ACC ↑	AUROC ↑	FPR@95 ↓	OSCR ↑	NMI ↑	ARI ↑
RepMix (ECCV 2022)	87.57 ± 7.84	70.57 ± 11.82	81.55 ± 6.43	66.93 ± 13.22	12.86 ± 10.66	11.22 ± 9.30
POSE (CVPR 2023)	71.83 ± 8.99	71.35 ± 5.97	88.16 ± 4.19	57.57 ± 9.72	2.76 ± 1.23	2.47 ± 0.94
ASV (Pattern Recogn. Lett. 2024)	57.03 ± 4.30	47.88 ± 3.97	91.97 ± 1.29	32.16 ± 3.44	6.66 ± 6.71	5.75 ± 5.75
Cioni et al. (ECCVW 2024)	<u>93.75</u> ± 7.40	77.45 ± 12.31	67.65 ± 20.53	75.92 ± 13.61	<u>60.93</u> ± 21.52	<u>57.91</u> ± 22.14
FSD (CVPR 2025)	82.85 ± 5.07	<u>82.95</u> ± 6.70	<u>44.85</u> ± 8.60	60.44 ± 15.55	13.87 ± 6.52	12.31 ± 6.49
LIDA (CVPR 2026)	87.06 ± 8.28	79.02 ± 6.74	59.00 ± 17.25	75.17 ± 7.96	14.42 ± 15.27	12.34 ± 13.28
IncreFA (CVPR 2026)	91.44 ± 7.97	80.79 ± 5.17	60.98 ± 17.33	<u>75.96</u> ± 5.58	24.76 ± 14.42	20.97 ± 12.39
RdH (Ours)	**98.64** ± 2.16	**92.84** ± 1.88	**28.43** ± 6.98	**84.50** ± 1.85	**79.15** ± 10.46	**78.90** ± 13.35

used to determine the threshold; unknown-source identities are used only to compute the clustering metrics.

Baselines and comparison settings. We compare with representative methods for open-set attribution, source verification, and recent AI-generated image attribution, including RepMix Bui et al. (2022), POSE Yang et al. (2023), ASV Abady et al. (2024), Cioni et al. Cioni et al. (2024), FSD Nguyen et al. (2025), LIDA Wang et al. (2026b), and IncreFA Qin et al. (2026). All trainable methods use the same known-source data and follow the same Cross-Era settings. For unknown-source clustering, the same K-means procedure is applied to each attribution representation without additional training (see Appendix D).

4.2 Cross-Era Open-Set Attribution

Table 1 reports the mean performance and cross-protocol standard deviation over S1–S5. RdH ranks first in all six macro-averaged metrics. Compared with the strongest competing average in each column, it improves ACC by 4.89 points, AUROC by 9.89 points, and OSCR by 8.54 percentage points, while reducing FPR@95 by 16.42 points. The gains are larger for unknown-source structure preservation: RdH improves NMI by 18.22 points and ARI by 20.99 points. These results show that the improvement is not confined to known-source classification or unknown rejection; the learned representation also retains substantially more information about the internal source structure of unknown-source evaluation samples. In contrast, the strongest baselines tend to specialize in different objectives: Cioni et al. provide the strongest competing ACC and clustering results, FSD gives the strongest competing AUROC and FPR@95, and IncreFA obtains the second-best OSCR.

The protocol-wise results in Appendix E.1 further show that RdH achieves the highest AUROC and lowest FPR@95 among the compared prior methods in each of the five settings. Its AUROC remains between 90.58% and 95.70%, and its OSCR remains between 81.24% and 86.36%, yielding cross-protocol standard deviations of only 1.88 and 1.85 points, respectively. This stability is notable because the protocols range from balanced and diffusion-dominant compositions to heterogeneous architectures, closely related lineages, and a U-Net→DiT shift. S4 remains particularly difficult: concentrating Janus-family generators among the unknown sources causes most competing methods to reach their worst or near-worst rejection results, whereas RdH retains 90.58% AUROC and 69.45% NMI. RdH does not obtain the best OSCR in every individual protocol, so its OSCR advantage should be interpreted as the strongest and most stable average trade-off across settings rather than uniform dominance in every case.

4.3 Evidence for Compositional Fingerprints

We first examine whether atoms learned from training generators can represent newer generators. For each unknown source, we compare its ten most active atoms with the highly active atoms of known sources. Across 20 protocol–unknown-source pairs, 8.4 of the ten atoms overlap on average, as shown in Figure 3(a). Thus, newer generators can largely reuse atoms learned from earlier generators rather than requiring an entirely separate feature set.

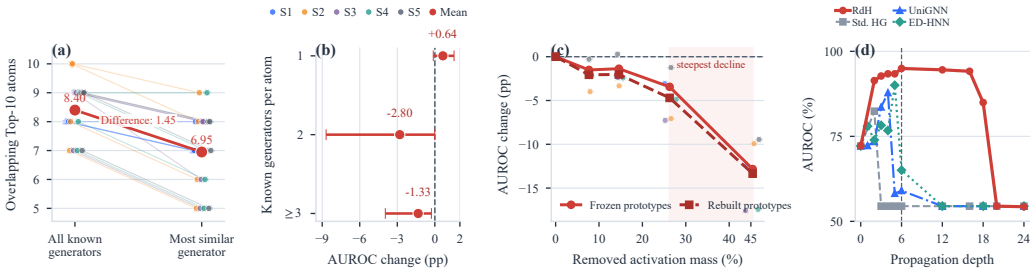


Figure 3: Evidence for atom combinations. Colored markers denote S1–S5, and red markers or curves denote their mean. (a) Top-10 atom overlap. (b) Masking atoms shared by different numbers of training generators. (c) Cumulative atom masking with frozen and rebuilt prototypes.(d) AUROC as a function of propagation depth.

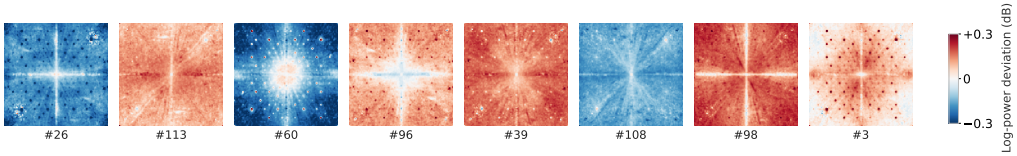


Figure 4: Atom-associated spectral deviations. Each tile shows the activation-weighted log-power deviation from a common baseline. All atoms share the same dB scale.

We then test whether such reuse contributes to unknown rejection through atom masking in Figure 3(b). Atoms are grouped by whether they are activated by one, two, or at least three known sources. Masking atoms used by only one source does not cause a consistent AUROC decrease, whereas masking atoms shared by multiple sources produces clearer degradation. This contrast indicates that unknown-source discrimination relies more strongly on reusable atoms than on features specific to a single known source.

Figure 3(c) further removes increasing amounts of atom activation to test whether source information is dominated by a few individual atoms. Performance changes only moderately under limited deletion but degrades more strongly as more activation is removed, indicating progressive disruption of the atom combination. Rebuilding source prototypes after deletion produces a similar trend, which rules out prototype-space mismatch as the main explanation for the degradation. Together, these results support source fingerprints formed by combinations of reusable atoms rather than by a few dominant atoms.

Figure 4 complements the quantitative evidence for atom reuse and combinations. The atoms exhibit different axis-aligned, diffuse, and localized periodic deviations, suggesting that the learned vocabulary is associated with heterogeneous low-level image statistics Frank et al. (2020). This post-hoc analysis indicates statistical associations rather than one-to-one correspondence with explicit generation mechanisms (see Appendix G.1) Chandrasegaran et al. (2021).

4.4 Validation of the RdH Design

The preceding experiments establish the role of reusable atoms and their combinations. We next examine the contributions of the main RdH components in Table 3, where DU, RC, and DW denote dictionary update, restricted connections, and degree-aware weighting, respectively. Removing restricted connections reduces AUROC from 92.84 to 85.52, consistent with their role in maintaining compact sample-specific atom combinations. Replacing degree-aware propagation with uniform hyperedge weighting reduces NMI from 79.15 to 59.32, supporting the suppression of broadly shared atoms during propagation. Fixing the atom dictionary also reduces AUROC to 86.31, showing that adapting the dictionary together with the dynamic incidence structure is important. The complete design achieves the best result across all four metrics.

Table 2 evaluates RdH with four visual backbones. CLIP provides the strongest overall results, while the method can also be instantiated with DINOv2, SigLIP, and ResNet-50. The performance differences indicate that the quality of the frozen visual representation remains important to open-set attribution and unknown-source structure preservation (see Appendix Sections E.2–E.3).

Table 2: Backbone sensitivity on Cross-Era. Mean ± std. across five distribution shifts.

Backbone	AUROC ↑	FPR@95 ↓	OSCR ↑	NMI ↑
CLIP	**92.84** ± 1.88	**28.43** ± 6.98	**84.50** ± 1.85	**79.15** ± 10.46
DINOv2	86.62 ± 9.17	35.72 ± 20.16	84.19 ± 11.42	56.95 ± 17.74
SigLIP	79.05 ± 7.45	55.66 ± 12.81	76.68 ± 9.21	29.61 ± 13.34
ResNet-50	75.25 ± 9.16	58.63 ± 11.36	71.64 ± 10.51	31.87 ± 15.63

Table 3: Component ablation on Cross-Era. Results are reported as mean ± std. over the five distribution shifts.

DU	RC	DW	AUROC ↑	FPR@95 ↓	OSCR ↑	NMI ↑
✓		✓	85.52 ± 12.98	42.07 ± 22.04	77.74 ± 10.45	74.29 ± 14.05
✓	✓		86.15 ± 4.03	57.17 ± 14.14	77.64 ± 4.10	59.32 ± 14.52
	✓	✓	86.31 ± 7.14	46.27 ± 13.37	78.82 ± 5.67	75.59 ± 13.17
✓	✓	✓	**92.84** ± 1.88	**28.43** ± 6.98	**84.50** ± 1.85	**79.15** ± 10.46

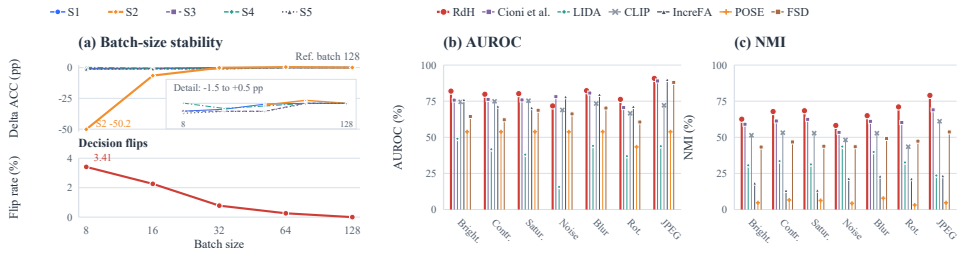


Figure 5: Inference stability and input robustness.(a) Sensitivity to test batch construction. (b–c)For each corruption condition, 50% of the test images are perturbed. AUROC and NMI under seven image changes on S1.

4.5 Inference Stability and Robustness

Because RdH constructs the incidence structure from the current test samples, its predictions may depend on batch size and composition. We therefore regroup the same 180 images under different batch sizes and known-to-unknown ratios while keeping all learned parameters fixed. As shown in Figure 5(a), the decision-flip rate decreases from 3.41% at a batch size of 8 to zero at 128, while known-source ACC remains stable outside the smallest batches. The main exception is S2, whose ACC drops by 50.20 pp at batch size 8. Since S2 is dominated by diffusion-based known sources, their more similar generative characteristics make the incidence structure more sensitive to limited relational context. This degradation quickly diminishes as the batch size increases. Thus, the observed batch dependence is mainly concentrated in small-batch settings (see Appendix H.2).

Practical forensic images may also undergo common changes such as compression, blur, and noise. We therefore evaluate seven moderate image changes on S1. Figure 5(b–c) shows that RdH achieves the highest AUROC in six conditions and the highest NMI in all seven, while additive noise remains the main exception. Under additive noise, RdH obtains 71.74% AUROC compared with 78.16% for Cioni et al. These results show that the current method remains sensitive to small test batches and additive noise (see Appendix H.3).

5 Conclusion

Open-set AI-generated image attribution is challenging because new generators may share forensic traces with generators in the known-source set. We introduce compositional source fingerprints and instantiate them with RdH, which represents each image through combinations of reusable fingerprint atoms. Experiments across five Cross-Era settings support our central hypothesis: modeling source fingerprints compositionally benefits both known-source attribution and unknown rejection, while preserving structure among unknown sources. Atom reuse and intervention analyses further show that attribution relies on combinations of reusable atoms rather than a few dominant patterns.

RdH still cannot fully attribute previously unseen generators without source-specific evidence. A key limitation is that learned atoms remain latent: their correspondence to concrete generation mechanisms is unclear, and correlations among atoms are not fully disentangled. In addition, the current inference procedure constructs the incidence structure from companion test samples, making it more sensitive to extremely small batches where relational evidence is limited. Future work will therefore focus on linking atoms to identifiable generative mechanisms, separating correlated responses, and developing batch-independent or reference-assisted incidence construction for single-image inference. These directions aim to improve both the transferability of compositional fingerprints to genuinely novel generators and the practicality of RdH for real world deployment.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing and improving the readability of the manuscript. Additionally, we used generative AI tools to provide suggestions for the presentation of selected tables and figures. We reviewed and verified all AI-assisted work. We take responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.

REFERENCES

Lydia Abady, Jun Wang, Benedetta Tondi, and Mauro Barni. A siamese-based verification system for open-set architecture attribution of synthetic images. *Pattern Recognition Letters*, 180:75–81, 2024.

Abhijit Bendale and Terrance E. Boult. Towards open set deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1563–1572, 2016.

Tu Bui, Ning Yu, and John Collomosse. RepMix: Representation mixing for robust attribution of synthesized images. In *European Conference on Computer Vision*, pp. 146–163, 2022.

George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. FakeInversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10759–10769, 2024.

Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung. A closer look at fourier spectrum discrepancies for CNN-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7200–7209, 2021.

Guanzi Chen, Jiying Zhang, Xi Xiao, and Yang Li. Preventing over-smoothing for hypergraph neural networks. *arXiv preprint arXiv:2203.17159*, 2022.

Dario Cioni, Christos Tzelepis, Lorenzo Seidenari, and Ioannis Patras. Are CLIP features all you need for universal synthetic image origin attribution? In *European Conference on Computer Vision Workshops*, pp. 363–382, 2024.

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

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, pp. 8780–8794, 2021.

Yuzhen Ding, Nupur Thakur, and Baoxin Li. Does a GAN leave distinct model-specific fingerprints? In *Proceedings of the British Machine Vision Conference*, 2021. doi: 10.5244/C.35.53.

Ricard Durall, Margret Keuper, and Janis Keuper. 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*, pp. 7890–7899, 2020.

Shengbang Fang, Tai D. Nguyen, and Matthew C. Stamm. Open set synthetic image source attribution. In *Proceedings of the British Machine Vision Conference*. BMVA, 2023. URL `https://papers.bmvc2023.org/0659.pdf`.

Jianwei Fei, Yunshu Dai, Peipeng Yu, Zhe Kong, Jiantao Zhou, and Zhihua Xia. Scalable dual fingerprinting for hierarchical attribution of text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15025–15034, 2025.

Jianwei Fei, Yunshu Dai, Peipeng Yu, Zhihua Xia, Dasara Shullani, Daniele Baracchi, and Alessandro Piva. One for all: Synthesis-free fingerprint learning for attribution of in-the-wild synthetic images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35419–35427, 2026. doi: 10.1609/aaai.v40i42.40851.

Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3558–3565, 2019.

Yifan Feng, Jiangang Huang, Shaoyi Du, Shihui Ying, Jun-Hai Yong, Yipeng Li, Guiguang Ding, Rongrong Ji, and Yue Gao. Hyper-YOLO: When visual object detection meets hypergraph computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2388–2401, 2025.

Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, and Teddy Furon. The stable signature: Rooting watermarks in latent diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22466–22477, 2023.

Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9284–9292, 2021.

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

Yue Gao, Zizhao Zhang, Haojie Lin, Xibin Zhao, Shaoyi Du, and Changqing Zou. Hypergraph learning: Methods and practices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2548–2566, 2022.

Sharath Girish, Saksham Suri, Sai Saketh Rambhatla, and Abhinav Shrivastava. Towards discovery and attribution of open-world GAN generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14094–14103, 2021.

Sam Gunn, Xuandong Zhao, and Dawn Song. An undetectable watermark for generative image models. In *Proceedings of the International Conference on Learning Representations*, 2025.

Kai Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8400–8408, 2019.

Jing Huang and Jie Yang. UniGNN: A unified framework for graph and hypergraph neural networks. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 2563–2569, 2021.

Jianwen Jiang, Yuxuan Wei, Yifan Feng, Jingxuan Cao, and Yue Gao. Dynamic hypergraph neural networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 2635–2641, 2019. doi: 10.24963/ijcai.2019/366.

Mengqi Lei, Yihong Wu, Siqi Li, Xinhu Zheng, Juan Wang, Yue Gao, and Shaoyi Du. SoftHGNN: Soft hypergraph neural networks for general visual recognition. *arXiv preprint arXiv:2505.15325*, 2025.

Ming Li, Yuchen Fang, Yifei Wang, Hao Feng, Yifan Gu, Lu Bai, and Pietro Liò. Deep hypergraph neural networks with tight framelets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 18385–18392, 2025.

Chi Liu, Tianqing Zhu, Yuan Zhao, Jun Zhang, and Wanlei Zhou. Disentangling different levels of GAN fingerprints for task-specific forensics. *Computer Standards & Interfaces*, 89:103825, 2024a. doi: 10.1016/j.csi.2023.103825.

Fengyuan Liu, Haochen Luo, Yiming Li, Philip Torr, and Jindong Gu. Which model generated this image? a model-agnostic approach for origin attribution. In *European Conference on Computer Vision*, pp. 282–301, 2024b.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do GANs leave artificial fingerprints? In *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval*, pp. 506–511, 2019.

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

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

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

Haotian Qin, Dongliang Chang, Yueying Gao, Yuexuan Tan, Lei Chen, and Zhanyu Ma. IncreFA: Breaking the static wall of generative model attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35405–35415, 2026.

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

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3418–3432, 2023.

Hae Jin Song and Laurent Itti. Riemannian-geometric fingerprints of generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11425–11435, 2025.

Hae Jin Song, Mahyar Khayatkhoei, and Wael AbdAlmageed. ManiFPT: Defining and analyzing fingerprints of generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10791–10801, 2024.

Chao Wang, Zijin Yang, Yaofei Wang, Weiming Zhang, and Kejiang Chen. AEDR: Training-free AI-generated image attribution via autoencoder double-reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9675–9683, 2026a.

Hongsong Wang, Renxi Cheng, Chaolei Han, and Jie Gui. Attribution as retrieval: Model-agnostic AI-generated image attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14062–14072, 2026b.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8695–8704, 2020.

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

Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Shiyu Wu, Shuyan Li, Jing Li, Jing Liu, and Yequan Wang. Few-shot synthetic image attribution: Identifying unseen generators with limited samples. In *European Conference on Computer Vision*, 2026.

Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Detecting origin attribution for text-to-image diffusion models. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 8764–8774, 2025a.

Rui Xu, Mengya Hu, Deren Lei, Yaxi Li, David Lowe, Alex Gorevski, Mingyu Wang, Emily Ching, and Alex Deng. InvisMark: Invisible and robust watermarking for AI-generated image provenance. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 909–918, 2025b.

Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. HyperGCN: A new method for training graph convolutional networks on hypergraphs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Tianyun Yang, Danding Wang, Fan Tang, Xinying Zhao, Juan Cao, and Sheng Tang. Progressive open space expansion for open-set model attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15856–15865, 2023.

Tianyun Yang, Danding Wang, Juan Cao, and Chang Xu. Model synthesis for zero-shot model attribution. *IEEE Transactions on Multimedia*, 27:8967–8980, 2025. doi: 10.1109/TMM.2025.3607778.

Zijin Yang, Kai Zeng, Kejiang Chen, Han Fang, Weiming Zhang, and Nenghai Yu. Gaussian shading: Provable performance-lossless image watermarking for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12162–12171, 2024.

Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7555–7565, 2019.

Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14448–14457, 2021.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. In *Proceedings of the IEEE International Workshop on Information Forensics and Security*, pp. 1–6, 2019.

Yu Zheng, Boyang Gong, Fanye Kong, Yueqi Duan, Bingyao Yu, Wenzhao Zheng, Lei Chen, Jiwen Lu, and Jie Zhou. Learning counterfactually decoupled attention for open-world model attribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 122–132, 2025.

Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10867–10875, 2021.

Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning with hypergraphs: Clustering, classification, and embedding. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Mengyao Zhou, Zhiheng Zhou, Xiao Han, Xingqin Qi, Guanghui Wang, and Guiying Yan. Tackling over-smoothing on hypergraphs: A ricci flow-guided neural diffusion approach. *arXiv preprint arXiv:2603.15696*, 2026.

Peng Zhou, Zongqian Wu, Xiangxiang Zeng, Guoqiu Wen, Junbo Ma, and Xiaofeng Zhu. Totally dynamic hypergraph neural networks. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 2476–2483, 2023. doi: 10.24963/ijcai.2023/275.

Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. AIGI-Holmes: Towards explainable and generalizable AI-generated image detection via multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18746–18758, 2025.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. GenImage: A million-scale benchmark for detecting AI-generated image. In *Advances in Neural Information Processing Systems*, 2023.

APPENDIX

Organization. The appendix provides formal analysis, method and optimization details, evaluation protocols, implementation settings, protocol-wise results, and additional analyses of atom combinations, propagation depth, test-batch sensitivity, and input corruptions.

A FORMAL ANALYSIS OF RESTRICTED AND DEGREE-AWARE PROPAGATION

This section analyzes the incidence- and degree-dependent coupling of the linear pre-activation propagation matrix in Eq. (A.50). The assumptions and bounds are stated below.

A.1 NOTATION AND SCOPE OF THE ANALYSIS

Let

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{H}) \tag{A.1}$$

denote an image–atom hypergraph, where $\mathcal{V} = \{v_1, \ldots, v_N\}$ is the set of image nodes, $\mathcal{E} = \{e_1, \ldots, e_M\}$ is the set of atom-induced hyperedges, and

$$\mathbf{H} \in \mathbb{R}_{\geq 0}^{N \times M} \tag{A.2}$$

is the restricted weighted incidence matrix.

The support of node i is

$$\mathcal{N}_H(i) = \{j : H_{ij} > 0\}. \tag{A.3}$$

Write $s_i = |\mathcal{N}_H(i)|$. Because the same row-wise Top-K_H rule is applied during training and inference,

$$s_i \leq K_H \leq M. \tag{A.4}$$

The weighted node and hyperedge degrees are

$$d_{v,i} = \sum_{j=1}^{M} H_{ij}, \qquad d_{e,j} = \sum_{i=1}^{N} H_{ij}. \tag{A.5}$$

As in Eq. (A.48)–(A.49), we use clamped degrees in the propagation operator:

$$\begin{aligned} \mathbf{D}_v &= \mathrm{diag}(\widehat{d}_{v,1}, \ldots, \widehat{d}_{v,N}), \qquad \widehat{d}_{v,i} = \max(d_{v,i}, 10^{-6}), \\ \mathbf{D}_e &= \mathrm{diag}(\widehat{d}_{e,1}, \ldots, \widehat{d}_{e,M}), \qquad \widehat{d}_{e,j} = \max(d_{e,j}, 10^{-6}). \end{aligned} \tag{A.6}$$

The nonnegative pre-activation propagation matrix is

$$\mathbf{T}_\beta = \mathbf{D}_v^{-\frac{1}{2}} \mathbf{H} \mathbf{D}_e^{-(1+\beta)} \mathbf{H}^\top \mathbf{D}_v^{-\frac{1}{2}}, \tag{A.7}$$

where $\beta > 0$ is the degree-aware exponent used in the model. For comparison only, setting $\beta = 0$ recovers the corresponding inverse-degree-normalized operator without the additional attenuation factor. The entry

$$(\mathbf{T}_\beta)_{ik} = \frac{1}{\sqrt{\widehat{d}_{v,i} \widehat{d}_{v,k}}} \sum_{j=1}^{M} H_{ij} H_{kj} \widehat{d}_{e,j}^{-(1+\beta)} \tag{A.8}$$

measures the normalized coupling between image nodes i and k through their shared atoms.

For analysis only, let the image nodes be partitioned according to their underlying sources:

$$\mathcal{V} = \bigcup_{c=1}^{C^\star} \mathcal{V}_c, \qquad \mathcal{V}_c \cap \mathcal{V}_{c'} = \varnothing \quad (c \neq c'). \tag{A.9}$$

This partition may include both known and unknown sources. It is used only to state and analyze source-aligned propagation behavior and is not required by the incidence construction.

For a source $\mathcal{V}_c$ and an atom e_j, define the weighted incidence mass from inside and outside the source as

$$q^+_{j,c} = \sum_{i \in \mathcal{V}_c} H_{ij}, \qquad q^-_{j,c} = \sum_{k \notin \mathcal{V}_c} H_{kj}. \tag{A.10}$$

Their sum is the raw hyperedge degree:

$$d_{e,j} = q^+_{j,c} + q^-_{j,c}. \tag{A.11}$$

For $d_{e,j} > 0$, we define the cross-source incidence fraction

$$\rho_{j,c} = \frac{q^-_{j,c}}{d_{e,j}}. \tag{A.12}$$

For $d_{e,j} = 0$, set $\rho_{j,c} = 0$; such a column contributes nothing to propagation. The quantity $\rho_{j,c}$ measures how much of the weighted incidence mass of an active atom j lies outside source c. This definition does not assign atom j exclusively to source c; the same atom may be reused by multiple sources.

We use the following boundedness conditions.

Assumption A.1 (Bounded Incidence and Feature Norms). There exist constants $h_{\max} > 0$, $x_{\max} > 0$, $d_{v,\min} > 0$, and $d_{e,\min} > 0$ such that

$$0 \le H_{ij} \le h_{\max}, \qquad \|\mathbf{x}_i\|_2 \le x_{\max}, \tag{A.13}$$

and every active node and hyperedge has effective degrees satisfying

$$\widehat{d}_{v,i} \ge d_{v,\min}, \qquad \widehat{d}_{e,j} \ge d_{e,\min}. \tag{A.14}$$

Assumption A.2 (Conditionally Bounded Cross-Source Co-activation). For an image node $i \in \mathcal{V}_c$ and each retained atom $j \in \mathcal{N}_H(i)$, the cross-source incidence fraction satisfies

$$\mathbb{E}[\rho_{j,c} \mid \mathbf{H}_{i,:}] \le \varepsilon, \tag{A.15}$$

where $\varepsilon \in [0, 1]$ and the expectation is over batch/incidence realizations conditional on node i's incidence row and source membership. This condition is applied only when $H_{ij} > 0$.

Assumption A.2 states that the atoms strongly selected by an image are not dominated, in expectation, by incidence mass from other sources. It does not require perfectly source-private atoms and remains compatible with atom reuse across generators.

A.2 Restricted Incidence and Cross-Source Propagation

For a node $i \in \mathcal{V}_c$, define its one-step cross-source coupling mass as

$$\kappa^{(c)}_i = \sum_{k \notin \mathcal{V}_c} (\mathbf{T}_\beta)_{ik}. \tag{A.16}$$

This quantity measures the total linearized propagation weight received by node i from nodes belonging to other sources.

Proposition A.1 (Restricted Incidence Bounds Cross-Source Coupling). *Under Assumptions A.1 and A.2, the expected one-step cross-source coupling of any node $i \in \mathcal{V}_c$ satisfies*

$$\mathbb{E}\left[\kappa^{(c)}_i\right] \le \varepsilon\, d^{-\beta}_{e,\min} \sqrt{\frac{K_H h_{\max}}{d_{v,\min}}}. \tag{A.17}$$

The same bound applies during training and inference because both use the same Top-K_H restriction.

Proof. Using Eq. (A.8), we have

$$\begin{aligned} \kappa^{(c)}_i &= \sum_{k \notin \mathcal{V}_c} \frac{1}{\sqrt{\widehat{d}_{v,i}\widehat{d}_{v,k}}} \sum_{j=1}^{M} H_{ij} H_{kj} \widehat{d}^{-(1+\beta)}_{e,j} \\ &= \frac{1}{\sqrt{\widehat{d}_{v,i}}} \sum_{j \in \mathcal{N}_H(i)} H_{ij} \widehat{d}^{-(1+\beta)}_{e,j} \sum_{k \notin \mathcal{V}_c} \frac{H_{kj}}{\sqrt{\widehat{d}_{v,k}}}. \end{aligned} \tag{A.18}$$

By Assumption A.1, every node with $H_{kj} > 0$ has $\widehat{d}_{v,k} \geq d_{v,\min}$, and therefore

$$\begin{aligned}\kappa_i^{(c)} &\leq \frac{1}{\sqrt{\widehat{d}_{v,i} d_{v,\min}}} \sum_{j \in \mathcal{N}_H(i)} H_{ij} \widehat{d}_{e,j}^{-(1+\beta)} q_{j,c}^{-} \\ &= \frac{1}{\sqrt{\widehat{d}_{v,i} d_{v,\min}}} \sum_{j \in \mathcal{N}_H(i)} H_{ij} \widehat{d}_{e,j}^{-\beta} \frac{d_{e,j}}{\widehat{d}_{e,j}} \rho_{j,c} \\ &\leq \frac{1}{\sqrt{\widehat{d}_{v,i} d_{v,\min}}} \sum_{j \in \mathcal{N}_H(i)} H_{ij} \widehat{d}_{e,j}^{-\beta} \rho_{j,c}, \end{aligned} \tag{A.19}$$

where Eq. (A.12) is used in the second line and $d_{e,j} \leq \widehat{d}_{e,j}$ in the last line.

First using $\widehat{d}_{e,j} \geq d_{e,\min}$, then taking the expectation conditional on $\mathbf{H}_{i,:}$ and applying Assumption A.2 gives

$$\begin{aligned}\mathbb{E}\Big[\kappa_i^{(c)} \mid \mathbf{H}_{i,:}\Big] &\leq \frac{d_{e,\min}^{-\beta}}{\sqrt{\widehat{d}_{v,i} d_{v,\min}}} \sum_{j \in \mathcal{N}_H(i)} H_{ij} \mathbb{E}[\rho_{j,c} \mid \mathbf{H}_{i,:}] \\ &\leq \varepsilon d_{e,\min}^{-\beta} \frac{d_{v,i}}{\sqrt{\widehat{d}_{v,i} d_{v,\min}}}. \end{aligned} \tag{A.20}$$

The row-wise restriction and bounded incidence weights imply

$$d_{v,i} = \sum_{j \in \mathcal{N}_H(i)} H_{ij} \leq K_H h_{\max}. \tag{A.21}$$

Since $d_{v,i} \leq \widehat{d}_{v,i}$, taking the unconditional expectation yields

$$\begin{aligned}\mathbb{E}\Big[\kappa_i^{(c)}\Big] &\leq \frac{\varepsilon d_{e,\min}^{-\beta}}{\sqrt{d_{v,\min}}} \mathbb{E}\left[\frac{d_{v,i}}{\sqrt{\widehat{d}_{v,i}}}\right] \\ &\leq \varepsilon d_{e,\min}^{-\beta} \sqrt{\frac{K_H h_{\max}}{d_{v,\min}}}. \end{aligned} \tag{A.22}$$

This is Eq. (A.17). If $d_{v,i} = 0$, then the incidence row and $\kappa_i^{(c)}$ are zero, so the same bound holds. ■

Proposition A.1 also bounds the cross-source feature contribution before the learnable transformation. Define

$$\mathbf{r}_i^{(c)} = \widehat{d}_{v,i}^{-1/2} \sum_{j \in \mathcal{N}_H(i)} H_{ij} \widehat{d}_{e,j}^{-(1+\beta)} \left[\sum_{k \notin \mathcal{V}_c} H_{kj} \widehat{d}_{v,k}^{-1/2} \mathbf{x}_k\right] \tag{A.23}$$

as the other-source part of the aggregation. By the triangle inequality and Assumption A.1,

$$\begin{aligned}\mathbb{E}\Big[\|\mathbf{r}_i^{(c)}\|_2\Big] &\leq x_{\max} \mathbb{E}\Big[\kappa_i^{(c)}\Big] \\ &\leq \varepsilon x_{\max} d_{e,\min}^{-\beta} \sqrt{\frac{K_H h_{\max}}{d_{v,\min}}}. \end{aligned} \tag{A.24}$$

If the subsequent linear transformation $\mathbf{\Theta}^{(l)}$ has spectral norm at most $w_{\max}$, its transformed contribution is further bounded by $w_{\max}$ times the right-hand side of Eq. (A.24). The bound is pre-activation and does not claim a coordinatewise decrease after the nonlinear activation.

This result does not imply that all cross-source interactions should be removed. Shared atoms may represent reusable latent patterns that are valid across several generators. The support comparison is informative because the executed Top-K_H rule enforces $K_H < M$.

For comparison, an unrestricted dense row replaces K_H by M in Eq. (A.17). This does not assert that every dense hypergraph has larger cross-source coupling, because actual coupling also depends on the incidence values and leakage fractions.

A.3 Degree-Aware Attenuation of Irrelevant Aggregation

We next analyze the $\mathbf{D}_e^{-\beta}$ factor in the propagation operator. Consider an atom-induced hyperedge e_j and a reference source $\mathcal{V}_c$. Its message before redistribution to image nodes is

$$\mathbf{m}_j = \widehat{d}_{e,j}^{-(1+\beta)} \sum_{i=1}^{N} H_{ij} \widehat{d}_{v,i}^{-1/2} \mathbf{x}_i. \tag{A.25}$$

Using the source partition, this message can be decomposed as

$$\mathbf{m}_j = \mathbf{m}_{j,c}^{+} + \mathbf{m}_{j,c}^{-}, \tag{A.26}$$

where

$$\mathbf{m}_{j,c}^{+} = \widehat{d}_{e,j}^{-(1+\beta)} \sum_{i \in \mathcal{V}_c} H_{ij} \widehat{d}_{v,i}^{-1/2} \mathbf{x}_i \tag{A.27}$$

and

$$\mathbf{m}_{j,c}^{-} = \widehat{d}_{e,j}^{-(1+\beta)} \sum_{k \notin \mathcal{V}_c} H_{kj} \widehat{d}_{v,k}^{-1/2} \mathbf{x}_k. \tag{A.28}$$

The second term represents the contribution from nodes outside the reference source. It is called "irrelevant" only relative to the source-conditioned decomposition used in this analysis; the same nodes may provide useful evidence for their own sources.

Proposition A.2 (Degree-Aware Attenuation Bounds Irrelevant Hyperedge Contributions). *Under Assumption A.1, the irrelevant component of hyperedge e_j relative to source $\mathcal{V}_c$ satisfies*

$$\left\|\mathbf{m}_{j,c}^{-}\right\|_2 \le \frac{\rho_{j,c} x_{\max}}{\sqrt{d_{v,\min}}} \widehat{d}_{e,j}^{-\beta}. \tag{A.29}$$

Moreover, if $\rho_{j,c} \le \rho_{\max}$ for every atom retained by node $i \in \mathcal{V}_c$, then the total irrelevant message delivered to node i is bounded by

$$\left\|\mathbf{r}_{i,c}^{-}\right\|_2 \le \rho_{\max} x_{\max} d_{e,\min}^{-\beta} \sqrt{\frac{K_H h_{\max}}{d_{v,\min}}}, \tag{A.30}$$

where

$$\mathbf{r}_{i,c}^{-} = \widehat{d}_{v,i}^{-\frac{1}{2}} \sum_{j \in \mathcal{N}_H(i)} H_{ij} \mathbf{m}_{j,c}^{-}. \tag{A.31}$$

Proof. Using Eq. (A.28) and the triangle inequality,

$$\begin{aligned} \left\|\mathbf{m}_{j,c}^{-}\right\|_2 &= \left\| \widehat{d}_{e,j}^{-(1+\beta)} \sum_{k \notin \mathcal{V}_c} H_{kj} \widehat{d}_{v,k}^{-1/2} \mathbf{x}_k \right\|_2 \\ &\le \widehat{d}_{e,j}^{-(1+\beta)} \sum_{k \notin \mathcal{V}_c} H_{kj} \widehat{d}_{v,k}^{-1/2} \|\mathbf{x}_k\|_2. \end{aligned} \tag{A.32}$$

By Assumption A.1, each contributing node has $\|\mathbf{x}_k\|_2 \le x_{\max}$ and $\widehat{d}_{v,k} \ge d_{v,\min}$. Since $q_{j,c}^{-} = d_{e,j}\rho_{j,c} \le \widehat{d}_{e,j}\rho_{j,c}$,

$$\begin{aligned} \left\|\mathbf{m}_{j,c}^{-}\right\|_2 &\le \frac{x_{\max}}{\sqrt{d_{v,\min}}} \widehat{d}_{e,j}^{-(1+\beta)} q_{j,c}^{-} \\ &\le \frac{\rho_{j,c} x_{\max}}{\sqrt{d_{v,\min}}} \widehat{d}_{e,j}^{-\beta}, \end{aligned} \tag{A.33}$$

which proves Eq. (A.29).

For node $i \in \mathcal{V}_c$, applying the triangle inequality again gives

$$\begin{aligned} \left\|\mathbf{r}_{i,c}^{-}\right\|_2 &\le \widehat{d}_{v,i}^{-1/2} \sum_{j \in \mathcal{N}_H(i)} H_{ij} \|\mathbf{m}_{j,c}^{-}\|_2 \\ &\le \frac{x_{\max}}{\sqrt{\widehat{d}_{v,i} d_{v,\min}}} \sum_{j \in \mathcal{N}_H(i)} H_{ij} \rho_{j,c} \widehat{d}_{e,j}^{-\beta}. \end{aligned} \tag{A.34}$$

Using $\rho_{j,c} \le \rho_{\max}$, $\widehat{d}_{e,j} \ge d_{e,\min}$, $d_{v,i} \le K_H h_{\max}$, and $d_{v,i} \le \widehat{d}_{v,i}$, we obtain

$$\begin{aligned}\left\|\mathbf{r}_{i,c}^{-}\right\|_2 &\le \frac{\rho_{\max} x_{\max} d_{e,\min}^{-\beta}}{\sqrt{d_{v,\min}}} \frac{d_{v,i}}{\sqrt{\widehat{d}_{v,i}}} \\ &\le \rho_{\max} x_{\max} d_{e,\min}^{-\beta} \sqrt{\frac{K_H h_{\max}}{d_{v,\min}}}. \end{aligned} \tag{A.35}$$

This proves Eq. (A.30). ■

Corollary A.1 (Degree-Dependent Change in the Propagation Bound). For a fixed incidence matrix, the irrelevant-component bound at exponent $\beta > 0$ relative to $\beta = 0$ changes by the factor

$$\widehat{d}_{e,j}^{-\beta}. \tag{A.36}$$

This factor is smaller than one only for $\widehat{d}_{e,j} > 1$; it equals one at degree one and exceeds one for $0 < \widehat{d}_{e,j} < 1$.

Corollary A.1 characterizes the additional degree factor in the pre-activation propagation operator. For effective degrees above one, it reduces the stated upper bound; the claim does not apply to degrees below one and does not imply a coordinatewise decrease after the learnable transformation and nonlinear activation.

The proposition bounds the magnitude of other-source aggregation, not recognition performance. The same degree factor also changes source-consistent evidence.

B Complete Method and Optimization Details

This section details the forward computation and training objective. Training uses known-source samples only. The CLIP visual encoder and intermediate-feature projector are frozen; the feature-fusion gate, incidence parameters, hypergraph layers, classifier, and atom dictionary are trained or updated on known-source data.

B.1 Restricted Incidence Construction

Image and atom representations. For a mini-batch of B images, the frozen CLIP encoder and frozen intermediate-feature projector produce normalized mid-level and final features $\mathbf{f}_{i,\mathrm{mid}}, \mathbf{f}_{i,\mathrm{high}} \in \mathbb{R}^d$. A learnable gate $\boldsymbol{\alpha}_i \in (0,1)^d$ fuses them into image-node features $\mathbf{x}_i = \boldsymbol{\alpha}_i \odot \mathbf{f}_{i,\mathrm{mid}} + (\mathbf{1} - \boldsymbol{\alpha}_i) \odot \mathbf{f}_{i,\mathrm{high}}$, with $\boldsymbol{\alpha}_i = \sigma(\mathrm{MLP}_{\mathrm{gate}}([\mathbf{f}_{i,\mathrm{mid}}; \mathbf{f}_{i,\mathrm{high}}]))$. We collect these features as

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_B]^\top \in \mathbb{R}^{B \times d}. \tag{A.37}$$

The shared atom dictionary is

$$\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_M]^\top \in \mathbb{R}^{M \times d}, \tag{A.38}$$

where each atom $\mathbf{c}_j$ is a learned latent pattern and induces the corresponding hyperedge e_j. The atoms are data-driven representation elements and are not assumed to correspond one-to-one with explicit generator modules or generation operators.

For incidence construction, we normalize the image and atom features:

$$\bar{\mathbf{x}}_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|_2 + \epsilon}, \qquad \bar{\mathbf{c}}_j = \frac{\mathbf{c}_j}{\|\mathbf{c}_j\|_2 + \epsilon}. \tag{A.39}$$

The preliminary soft response of image i to atom j is

$$\widetilde{H}_{ij} = \frac{\exp(\bar{\mathbf{x}}_i^\top \bar{\mathbf{c}}_j / \tau)}{\sum_{m=1}^{M} \exp(\bar{\mathbf{x}}_i^\top \bar{\mathbf{c}}_m / \tau)}, \tag{A.40}$$

where $\tau > 0$ is the temperature. A sample-dependent threshold is predicted independently for each image:

$$\theta_i = f_\theta(\bar{\mathbf{x}}_i), \qquad \boldsymbol{\theta} = [\theta_1, \ldots, \theta_B]^\top. \tag{A.41}$$

The thresholded response is

$$R_{ij} = [\widetilde{H}_{ij} - \theta_i]_+, \tag{A.42}$$

where $[\cdot]_+$ denotes ReLU.

Row-wise support restriction. Both training and inference apply the same row-wise Top-K_H restriction. The implementation uses `torch.topk`; boundary ties therefore follow the active backend rather than an explicit stable ordering. For the formal properties below, we condition on a fixed deterministic tie resolution and define

$$S_{ij} = \begin{cases} 1, & j \in \mathrm{TopK}_{K_H}(\mathbf{R}_{i,:}), \\ 0, & \text{otherwise}, \end{cases} \tag{A.43}$$

and construct

$$\mathbf{H} = \mathbf{R} \odot \mathbf{S} = \left[\widetilde{\mathbf{H}} - \boldsymbol{\theta}\mathbf{1}_M^\top\right]_+ \odot \mathbf{S}. \tag{A.44}$$

Thus every image is incident to at most K_H atom-induced hyperedges in both training and inference. The restriction acts on support only; positive incidence values are not binarized.

The compositional feature is

$$\mathbf{h}_i = \frac{\mathbf{H}_{i,:}}{\|\mathbf{H}_{i,:}\|_2 + \epsilon}, \qquad \epsilon > 0, \tag{A.45}$$

and

$$\mathbf{H}_{\mathrm{norm}} = [\mathbf{h}_1, \ldots, \mathbf{h}_B]^\top. \tag{A.46}$$

A zero incidence row remains zero. Once the encoder, dictionary, threshold predictor, temperature, and Top-K_H tie rule are fixed, $\mathbf{h}_i$ depends only on image i.

Degree-aware propagation. The weighted node and hyperedge degrees are

$$d_{v,i} = \sum_{j=1}^{M} H_{ij}, \qquad d_{e,j} = \sum_{i=1}^{B} H_{ij}, \tag{A.47}$$

with positively lower-bounded matrices

$$\mathbf{D}_v = \mathrm{diag}(\widehat{d}_{v,1}, \ldots, \widehat{d}_{v,B}), \qquad \widehat{d}_{v,i} = \max(d_{v,i}, \epsilon), \tag{A.48}$$

$$\mathbf{D}_e = \mathrm{diag}(\widehat{d}_{e,1}, \ldots, \widehat{d}_{e,M}), \qquad \widehat{d}_{e,j} = \max(d_{e,j}, \epsilon). \tag{A.49}$$

For an L-layer model, let $\mathbf{X}^{(0)} = \mathbf{X}$. Layer l performs

$$\mathbf{X}^{(l)} = \phi_l\left(\mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{D}_e^{-(1+\beta)}\mathbf{H}^\top\mathbf{D}_v^{-1/2}\mathbf{X}^{(l-1)}\boldsymbol{\Theta}^{(l)}\right), \qquad \beta > 0, \tag{A.50}$$

where $\boldsymbol{\Theta}^{(l)}$ is learnable and ϕ_l is the layer activation. The final propagated representation is

$$\mathbf{X}' = \mathbf{X}^{(L)}. \tag{A.51}$$

Unlike $\mathbf{h}_i$, the propagated feature $\mathbf{X}'_i$ depends on companion images through hyperedge degrees and node–hyperedge–node aggregation.

B.2 Elementary Properties of the Atom Combination and Degree Weighting

Proposition B.1 (Preservation of the restricted atom combination). *Under the fixed-model and deterministic-tie assumptions above,*

$$\mathrm{supp}(\mathbf{h}_i) = \mathrm{supp}(\mathbf{H}_{i,:}), \qquad |\mathrm{supp}(\mathbf{h}_i)| \le K_H. \tag{A.52}$$

For any active atoms j and l,

$$\frac{h_{ij}}{h_{il}} = \frac{H_{ij}}{H_{il}}. \tag{A.53}$$

For any images i and k,

$$\mathbf{h}_i^\top \mathbf{h}_k = \sum_{j \in \mathrm{supp}(\mathbf{h}_i) \cap \mathrm{supp}(\mathbf{h}_k)} h_{ij} h_{kj}. \tag{A.54}$$

Consequently, $\mathbf{h}_i^\top \mathbf{h}_k > 0$ *if and only if the images share at least one active atom, and* $\mathbf{h}_i$ *is unchanged when other images in its inference batch are replaced or regrouped.*

Proof. Equation (A.44) applies ReLU and a row-wise Top-K_H mask, so $\mathbf{H}_{i,:}$ is nonnegative and has at most K_H positive coordinates. Equation (A.45) multiplies the row by one strictly positive scalar and therefore preserves its support and the ratio of positive coordinates. Expanding the inner product gives the stated intersection sum. The encoder, threshold predictor, and row-wise Top-K_H operation process image i independently under fixed parameters, so neither $\mathbf{H}_{i,:}$ nor $\mathbf{h}_i$ depends on companion images. If $\mathbf{H}_{i,:} = \mathbf{0}$, then $\mathbf{h}_i = \mathbf{0}$. ■

Proposition B.2 (Degree-dependent relative weighting of shared atoms). *Let*

$$\mathbf{P}_\beta = \mathbf{D}_v^{-1/2}\mathbf{H}\mathbf{D}_e^{-(1+\beta)}\mathbf{H}^\top\mathbf{D}_v^{-1/2}, \qquad \beta > 0. \tag{A.55}$$

The contribution of atom j *to images* i *and* k *is*

$$W_{ikj}^{(\beta)} = \frac{H_{ij}H_{kj}}{\sqrt{\widehat{d}_{v,i}\widehat{d}_{v,k}}\,\widehat{d}_{e,j}^{1+\beta}}, \qquad (\mathbf{P}_\beta)_{ik} = \sum_{j=1}^{M} W_{ikj}^{(\beta)}. \tag{A.56}$$

For positive contributions from atoms j *and* l *to the same pair,*

$$\frac{W_{ikj}^{(\beta)}/W_{ikl}^{(\beta)}}{W_{ikj}^{(0)}/W_{ikl}^{(0)}} = \left(\frac{\widehat{d}_{e,l}}{\widehat{d}_{e,j}}\right)^{\beta}. \tag{A.57}$$

If additional batch images leave the original incidence rows unchanged and strictly increase atom j*'s effective degree to* $\widehat{d}'_{e,j}$*, then*

$$\frac{W_{ikj}^{(\beta)\prime}}{W_{ikj}^{(\beta)}} = \left(\frac{\widehat{d}_{e,j}}{\widehat{d}'_{e,j}}\right)^{1+\beta} < 1 \tag{A.58}$$

for every fixed original pair with $H_{ij}H_{kj} > 0$*.*

Proof. Expanding Eq. (A.55) gives Eq. (A.56). The vertex-degree factor cancels when two atom contributions to the same pair are compared; division by the corresponding $\beta = 0$ ratio also cancels both incidence numerators, yielding Eq. (A.57). Proposition B.1 keeps the original rows and their vertex degrees fixed under regrouping. Substitution of the increased effective hyperedge degree yields Eq. (A.58). ■

These propositions concern the atom response and the coefficients before the nonlinear activation in Eq. (A.50). They do not imply coordinate-wise attenuation of the final feature, identify hyperedge degree with source selectivity, or assign atoms to physical generator components.

B.3 Dynamic Dictionary Update and Maintenance

The dictionary is learned only from known-source training data. Its training-time adaptation combines gradient-based optimization, an explicit incidence-weighted residual correction, and utility-based atom maintenance.

Incidence-weighted residual update. The dictionary-maintenance routine forms an atom-based reconstruction of the batch features:

$$\widehat{\mathbf{X}} = \mathbf{H}_{\mathrm{norm}}\mathbf{C}. \tag{A.59}$$

The corresponding feature residual is

$$\mathbf{R}_X = \mathbf{X} - \mathbf{H}_{\mathrm{norm}}\mathbf{C}. \tag{A.60}$$

After the differentiable update for the current mini-batch, the dictionary receives the explicit residual correction

$$\mathbf{C}^+ = \mathbf{C} + \eta\,\mathbf{H}_{\mathrm{norm}}^\top(\mathbf{X} - \mathbf{H}_{\mathrm{norm}}\mathbf{C}), \tag{A.61}$$

where η is the dictionary residual-update rate. The multiplication by $\mathbf{H}_{\text{norm}}^{\top}$ assigns each reconstruction residual according to the current incidence weights. Each corrected atom is then normalized:

$$\mathbf{c}_j \leftarrow \frac{\mathbf{c}_j^{+}}{\|\mathbf{c}_j^{+}\|_2 + \epsilon}. \tag{A.62}$$

The residual update does not assign atoms to source labels and does not assume that an atom uniquely represents one source.

Running atom utility. To identify underused atoms, the training forward pass maintains an exponential moving average of the current, post-threshold incidence values:

$$u_j^{(t)} = \rho_u u_j^{(t-1)} + (1-\rho_u)\frac{1}{B}\sum_{i=1}^{B} H_{ij}^{(t)}, \tag{A.63}$$

where $\rho_u \in [0,1)$ is the utility momentum. This statistic is computed from the effective training incidence, not from the preliminary softmax responses.

An atom is marked as inactive when

$$u_j^{(t)} < 0.001 \tag{A.64}$$

at a scheduled revival call. Revival is invoked at the first, every thirtieth, and final mini-batch of an epoch, at each epoch end, and additionally every fifth epoch.

Inactive-atom reactivation. The active revival set at a scheduled call is

$$\mathcal{J}_t = \{j : u_j^{(t)} < 0.001\}. \tag{A.65}$$

When no reference feature is supplied to the revival routine, an inactive atom is reset from Gaussian noise:

$$\mathbf{c}_j \leftarrow \text{RowNorm}(\boldsymbol{\xi}_j), \qquad \boldsymbol{\xi}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \qquad j \in \mathcal{J}_t. \tag{A.66}$$

Revival maintains the dictionary and introduces no additional supervised loss.

The utility statistics, residual updates, and atom reactivation are used during training only. At the end of training, the dictionary is frozen and no test image can modify the atom basis.

B.4 Training Objective

For a batch of known-source samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{B}$, the propagated representation is classified by

$$\mathbf{z}_i = g(\mathbf{X}_i'), \tag{A.67}$$

where g is the known-source classification head. The supervised classification loss is

$$\mathcal{L}_{\text{cls}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(z_{i,y_i})}{\sum_{c=1}^{C}\exp(z_{ic})}. \tag{A.68}$$

Structural constraint. The restricted incidence is regularized by

$$\mathcal{L}_{\text{str}} = \frac{1}{BM}\|\mathbf{H}\|_1 - \frac{\alpha}{B}\sum_{j=1}^{M}\frac{1}{\widehat{d}_{e,j}}\left\|\sum_{i=1}^{B}\frac{H_{ij}}{\sqrt{\widehat{d}_{v,i}}}\mathbf{x}_i\right\|_2^2, \tag{A.69}$$

where $\alpha \geq 0$ balances sparsity and within-hyperedge feature alignment. With

$$a_{ij} = \frac{H_{ij}}{\sqrt{\widehat{d}_{v,i}}}, \tag{A.70}$$

the squared aggregate expands as

$$\left\|\sum_i a_{ij}\mathbf{x}_i\right\|_2^2 = \sum_{i,k} a_{ij}a_{kj}\mathbf{x}_i^{\top}\mathbf{x}_k. \tag{A.71}$$

The first term promotes sparse incidence, while the second rewards aligned features connected by the same atom. The inverse effective hyperedge degree normalizes the reward according to the current weighted degree.

Dictionary orthogonality. Let the row-normalized atom matrix be

$$\bar{\mathbf{C}} = [\bar{\mathbf{c}}_1, \ldots, \bar{\mathbf{c}}_M]^\top \tag{A.72}$$

and define

$$\mathcal{L}_{\text{reg}} = \frac{1}{M^2} \left\| \bar{\mathbf{C}}\bar{\mathbf{C}}^\top - \mathbf{I}_M \right\|_F^2 . \tag{A.73}$$

This term discourages redundant atom directions. It does not impose semantic independence or guarantee an interpretable meaning for every individual atom.

The complete differentiable objective is

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{str}}\mathcal{L}_{\text{str}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{A.74}$$

The residual dictionary correction in Eq. (A.61) and utility-based maintenance are update operations rather than additional loss terms.

B.5 Source Prototype Construction

After training, the CLIP encoder, frozen intermediate projector, fusion gate, atom dictionary, threshold parameters, and propagation/classification parameters are fixed. We recompute the compositional feature $\mathbf{h}_i$ for every known-source training image using the fixed model.

For source c, let

$$\mathcal{D}_c = \{i : y_i = c\} \tag{A.75}$$

denote its training-image index set. Its mean atom-combination feature is

$$\bar{\boldsymbol{\mu}}_c = \frac{1}{|\mathcal{D}_c|} \sum_{i \in \mathcal{D}_c} \mathbf{h}_i. \tag{A.76}$$

We retain the K_p largest dimensions:

$$\widetilde{\boldsymbol{\mu}}_c = \mathrm{TopK}_{K_p}\left(\bar{\boldsymbol{\mu}}_c\right), \tag{A.77}$$

where TopK_{K_p} preserves the original positive values in the selected dimensions and sets the remaining dimensions to zero. The prototype is then normalized:

$$\boldsymbol{\mu}_c = \mathrm{RowNorm}(\widetilde{\boldsymbol{\mu}}_c). \tag{A.78}$$

The support size K_p controls the number of characteristic dimensions retained in a source-level prototype and is distinct from the per-image support size K_H, which is used in both training and inference. The resulting prototype bank is

$$\mathbf{P} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C]^\top \in \mathbb{R}^{C \times M}. \tag{A.79}$$

The prototype bank is fixed throughout validation and testing. No unknown-source image or unknown-source label is used to construct the prototypes.

B.6 Unknown-Score Composition

The propagated representation produces the candidate known-source prediction

$$\hat{y}_i = \arg\max_c z_{ic}. \tag{A.80}$$

RdH combines three complementary indicators of inconsistency with the known-source set.

Prototype distance. The compositional inconsistency with the candidate source is

$$u_{\text{proto}}(i) = 1 - \mathrm{RowNorm}(\mathbf{h}_i)^\top \mathrm{RowNorm}(\boldsymbol{\mu}_{\hat{y}_i}). \tag{A.81}$$

Energy uncertainty. The implemented energy uses unit logit temperature:

$$E_i = -\log \sum_{c=1}^{C} \exp(z_{ic}). \tag{A.82}$$

A fixed monotone sigmoid maps the energy to $(0, 1)$:

$$u_{\text{energy}}(\mathbf{z}_i) = \sigma(0.07 E_i), \tag{A.83}$$

MSP uncertainty. The maximum-softmax-probability uncertainty is

$$u_{\mathrm{MSP}}(\mathbf{z}_i) = 1 - \max_c \frac{\exp(z_{ic})}{\sum_{r=1}^{C} \exp(z_{ir})}. \tag{A.84}$$

Following the parameterization in the main paper, define

$$u_{\mathrm{cls}}(\mathbf{z}_i) = \rho u_{\mathrm{energy}}(\mathbf{z}_i) + (1-\rho) u_{\mathrm{MSP}}(\mathbf{z}_i), \qquad \rho \in [0,1], \tag{A.85}$$

and

$$s_i = w_p u_{\mathrm{proto}}(i) + (1-w_p) u_{\mathrm{cls}}(\mathbf{z}_i), \qquad w_p \in [0,1]. \tag{A.86}$$

This is equivalent to direct weights $w_e = (1-w_p)\rho$ and $w_m = (1-w_p)(1-\rho)$. The reported score uses

$$w_p = 0.45, \qquad \rho = \frac{7}{11}, \qquad (w_p, w_e, w_m) = (0.45, 0.35, 0.20). \tag{A.87}$$

Higher values of s_i indicate lower consistency with the candidate known source. AUROC and FPR@95 are computed from the score ordering or ROC/CCR–FPR curves and do not use a single operating threshold. If a separate binary known/unknown decision is required for an application, it may be defined at a specified threshold γ by

$$\hat{y}_i^{\mathrm{open}} = \begin{cases} \text{unknown}, & s_i > \gamma, \\ \hat{y}_i, & s_i \le \gamma. \end{cases} \tag{A.88}$$

The score weights and energy transform are fixed across the reported protocols. Threshold-free AUROC and FPR@95 use the score ordering. Any experiment that reports a binary known/unknown decision must additionally specify its operating threshold γ and the data and criterion used to select it. Test unknown-source labels are not used for prototype construction or score fitting.

The auxiliary NCD evaluation is separate from unknown-score thresholds. All unknown-source evaluation samples are clustered using their compositional features $\mathbf{h}_i$, independently of whether any application-specific operating threshold would reject each individual sample.

B.7 Training and Inference Algorithms

Algorithm A.1 summarizes the reported from-scratch training path. Algorithm A.2 summarizes score-producing inference after the dictionary and source prototypes have been fixed. A binary threshold decision is optional and is not part of the reported AUROC or FPR@95 calculation. The column labeled OSCR reports open-set accuracy at $\gamma = \big(\mathrm{median}(s_{\mathrm{known}}) + \mathrm{median}(s_{\mathrm{unknown}})\big)/2$. The medians are computed from known and unknown test samples. Accuracy counts known samples that are accepted and correctly classified, plus unknown samples that are rejected, divided by the total number of test samples. This is not an area-under-curve metric or accuracy at a validation-selected threshold.

The atom dictionary and source prototype bank remain unchanged during inference. The compositional feature $\mathbf{h}_i$ is sample-wise under the fixed dictionary, whereas the propagated classification feature is context-dependent because hypergraph degrees and message passing are computed from the current test batch.

C Datasets and Evaluation Protocols

C.1 Cross-Era Protocols

We construct five Cross-Era protocols using earlier sources from GenImage as known sources and recent generators from AIGI-Holmes as unknown sources. Each protocol contains five known sources, including real images and four generators, and four unknown sources that are absent from training. The protocols cover different generator-evolution shifts, as summarized in Table A.1.

The five protocols are not pooled into a single evaluation set. A separate model, atom dictionary, classifier, and source-prototype bank are trained for each protocol. Cross-Era results are reported as the macro average of the five protocol-level metrics.

Algorithm A.1 Training RdH

Require: Known-source training set $\mathcal{D}_{\mathrm{tr}}$; frozen CLIP feature extractor F; learnable fusion gate and threshold parameters; atom dictionary $\mathbf{C}$; six propagation layers and classifier g.
Ensure: Trained RdH parameters and source prototype bank $\mathbf{P}$.
Initialize $\mathbf{C}$ and atom utilities $\{u_j\}_{j=1}^{M}$.
for each training epoch **do**
  for each known-source mini-batch $\{(I_i, y_i)\}_{i=1}^{B}$ **do**
    Extract frozen mid/final features and fuse them through the learnable gate to obtain $\mathbf{x}_i$.
    Compute preliminary responses $\widetilde{\mathbf{H}}$ using Eq. (A.40).
    Predict training-mode thresholds using Eq. (A.41).
    Apply thresholding and the row-wise Top-K_H mask to obtain $\mathbf{H}$.
    Normalize the incidence rows using Eq. (A.45).
    Compute $\mathbf{D}_v$ and $\mathbf{D}_e$.
    Propagate node features through six degree-aware layers using Eq. (A.50).
    Compute logits $\mathbf{z}_i = g(\mathbf{X}'_i)$.
    Compute the classification, structural, and dictionary regularization terms.
    Update differentiable model parameters by minimizing Eq. (A.74).
    Apply the residual dictionary update and atom normalization in Eqs. (A.61)–(A.62).
    At scheduled calls, revive atoms satisfying Eq. (A.64); their utility was updated during the training forward pass.
  end for
end for
Freeze the visual encoder, dictionary, incidence predictor, propagation layers, and classifier.
Recompute known-source training compositional features.
Construct source prototypes using Eqs. (A.76)–(A.79).
return Trained model and $\mathbf{P}$.

Algorithm A.2 Open-Set Inference with RdH

Require: Test images $\{I_i\}_{i=1}^{B}$; fixed model and dictionary $\mathbf{C}$; fixed prototype bank $\mathbf{P}$; fixed score parameters; optional application threshold γ.
Ensure: Candidate known-source predictions and unknown scores; optional binary decisions.
Extract frozen mid/final features and form fused image features $\mathbf{x}_i$.
Compute $\widetilde{\mathbf{H}}$ from the fixed dictionary.
Predict the sample-adaptive thresholds.
Apply thresholding and the same row-wise Top-K_H rule to obtain $\mathbf{H}$.
Normalize the incidence rows to obtain $\mathbf{h}_i$ and $\mathbf{H}_{\mathrm{norm}}$.
Compute $\mathbf{D}_v$ and $\mathbf{D}_e$ from the current test batch.
Perform the six degree-aware propagation layers and obtain $\mathbf{X}'$.
Compute logits $\mathbf{z}_i = g(\mathbf{X}'_i)$.
Set $\hat{y}_i \leftarrow \arg\max_c z_{ic}$.
Compute prototype, energy, and MSP uncertainties.
Compute the composite unknown score s_i using Eq. (A.86).
if an application threshold γ is specified **then**
  for $i = 1$ to B **do**
    if $s_i > \gamma$ **then**
      $\hat{y}_i^{\mathrm{open}} \leftarrow$ unknown.
    else
      $\hat{y}_i^{\mathrm{open}} \leftarrow \hat{y}_i$.
    end if
  end for
end if
return $\{\hat{y}_i, s_i, \mathbf{h}_i\}_{i=1}^{B}$ and, if requested, $\{\hat{y}_i^{\mathrm{open}}\}_{i=1}^{B}$.

Table A.1: Source compositions of the five Cross-Era protocols. Each protocol is trained and evaluated independently. S5 uses the source annotation originally indexed as split 6.

Protocol	Known Sources	Unknown Sources
Balanced	real, ADM, BigGAN, GLIDE, Stable Diffusion v1.5	FLUX, Stable Diffusion 3.5-Large, PixArt-XL, Show-o
Diffusion-dominant Known Set	real, GLIDE, Stable Diffusion v1.5, ADM, VQDM	Janus-Pro-7B, VAR, LlamaGen, Infinity
Heterogeneous Mix	real, Midjourney, Wukong, StyleGAN3, GLIDE	Janus, Janus-Pro-1B, Show-o, PixArt-XL
Lineage	real, Stable Diffusion v1.5, BigGAN, Midjourney, Wukong	Janus, Janus-Pro-1B, Janus-Pro-7B, Show-o
U-Net→DiT	real, GLIDE, Stable Diffusion v1.5, ADM, VQDM	FLUX, Stable Diffusion 3.5-Large, PixArt-XL, Infinity

Table A.2: Sample counts for the five Cross-Era protocols. ID and OOD denote known-source and unknown-source images, respectively.

Protocol	Train ID	Val ID	Val OOD	Test ID	Test OOD
Balanced	35,000	7,500	10,000	7,500	10,000
Diff.-Known	35,000	7,500	10,000	7,500	10,000
Hetero. Mix	35,000	7,500	9,999	7,500	10,000
Lineage	35,000	7,500	9,999	7,500	10,000
U-Net→DiT	35,000	7,500	10,000	7,500	10,000

Unknown-source labels are not used during training or clustering. For the auxiliary NCD evaluation, the normalized compositional features of all unknown test images are clustered using K-means with $k = 4$. The source labels are used only afterward to compute NMI and ARI.

C.2 Dataset Statistics and Preprocessing

Known-source images are first deduplicated by path and balanced by downsampling each source to the smallest available source size, capped at 10,000 images per source. With random seed 42, the resulting 50,000 known-source images are divided into training, validation, and test sets using a 70%/15%/15% split.

Unknown-source images are sampled approximately uniformly across the four sources and divided equally between validation and test. The final sample counts are reported in Table A.2.

The known-source subsets are approximately balanced across the five labels, and the unknown subsets are approximately balanced across the four held-out generators. No path-level overlap is present among the training, validation, and test manifests within each protocol.

All images are decoded as RGB and processed using the official preprocessing pipeline of the frozen CLIP ViT-L/14 encoder, including resizing, center cropping to 224×224, tensor conversion, and CLIP normalization. The same preprocessing is used for training, validation, prototype construction, and testing. No additional input corruption is applied in the standard evaluation.

Training uses only samples with nonnegative known-source labels. Validation and test evaluation combine known-source images with their corresponding unknown-source subsets. For NCD, the compositional features are ℓ_2-normalized before applying K-means with $k = 4$, `n_init=20`, and random seed 42.

D Implementation and Reproducibility Details

This section reports the model, optimization, evaluation, and runtime settings.

D.1 Model and Optimization Settings

RdH uses a frozen CLIP ViT-L/14 visual encoder. We extract the class token from transformer block 12 (zero-based indexing) and the final visual representation. The intermediate token is projected

Table A.3: Effective model and optimization settings.

Setting	Value
Visual encoder	CLIP ViT-L/14
Intermediate block	12 (zero-based)
Fusion / hidden dimension	768 / 512
Known sources / atoms	5 / 128
Hypergraph layers	6
Incidence support K_H	12 (training and inference)
Dropout	0.1
Training / evaluation batch size	128 / 256
Data-loader workers	4
Training epochs	10
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Scheduler	CosineAnnealingLR
Scheduler $T_{\max}$ / $\eta_{\min}$	50 / 1×10^{-5}
Gradient clipping	ℓ_2 norm 1.0
Prototype support K_p	16 atoms
Unknown clustering	K-means, $k = 4$
K-means restarts / seed	20 / 42

from 1024 to 768 dimensions by a frozen linear layer, and the two normalized features are fused by a learnable gate before hypergraph construction.

Table A.3 lists the effective settings used for all five Cross-Era protocols. The GenImage experiments use a different incidence/prototype support pair, $K_H = 16$ and $K_p = 32$.

The effective training loss is the three-term objective in Eq. (A.74):

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{str}}\mathcal{L}_{\text{str}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{A.89}$$

Here $\mathcal{L}_{\text{str}}$ is the restricted-incidence structural term in Eq. (A.69), and $\mathcal{L}_{\text{reg}}$ is the normalized dictionary regularizer in Eq. (A.73). The residual dictionary correction and atom maintenance are update operations, not additional differentiable loss terms.

The best checkpoint is selected by the minimum total loss on known validation samples. Unknown validation samples do not contribute to checkpoint selection. After training, the prototype bank is built from the known-source training loader using the selected checkpoint.

The atom dictionary is additionally corrected after each optimizer step using the incidence-weighted reconstruction residual, with update rate 0.001; each corrected atom is then normalized as in Eq. (A.62). Atom utilities are updated during each training forward pass. Atom revival follows the schedule specified in Appendix B.3: the first, every thirtieth, and final mini-batch of an epoch, each epoch end, and additionally every fifth epoch. At these scheduled calls, atoms with utility below 0.001 are reinitialized.

Incidence sparsity. Top-K_H is applied after thresholding and ReLU and before incidence normalization in training and inference. Cross-Era uses $K_H = 12$ and $K_p = 16$; GenImage uses $K_H = 16$ and $K_p = 32$.

D.2 Hyperparameter Search

Hyperparameter selection is performed before test evaluation. Training settings are selected using known-source validation loss, while the open-set score uses one fixed composition across all protocols:

$$s = 0.45s_{\text{proto}} + 0.35s_{\text{energy}} + 0.20s_{\text{MSP}}. \tag{A.90}$$

No test-source labels are used for model selection or score configuration.

Table A.3 summarizes the training settings; Eq. (A.90) specifies the open-set score weights.

Table A.4: Baseline implementation configurations.

Method	Backbone and optimization	OOD score / checkpoint
RepMix	Adapted official ECCV 2022 code; 224 input; batch 32; Adam, lr 10^{-4}, weight decay 5×10^{-4}; 30 epochs; mixup $\beta = 0.4$.	Validation-loss checkpoint callback; evaluation uses `last.ckpt`.
POSE	Adapted official CVPR 2023 code; 128×128 DCT input; custom DCT CNN; batch 8; 10 epochs; Adam with encoder lr 10^{-4} and augmentation-network lr 10^{-2}.	Checkpoints saved at epochs 5 and 10.
ASV	ImageNet EfficientNet-B4; 380 input; 512-dimensional embedding; batch 32; Adam, lr 10^{-4}.	Cosine-prototype logits with temperature 20; MSP-based unknown score.
Cioni/ACFA	OpenAI CLIP ViT-B/16; 224 input; block-7 frozen feature; scikit-learn logistic regression; approximately 4,000 samples per class.	Negative maximum known-class classifier score; logistic C selected on validation data.
FSD	Local reproduction with eight learnable 11×11 filters; batch 16; 10 epochs; lr 10^{-3}.	Three-component diagonal GMM; 0.05 quantile threshold.
LIDA	Adapted CVPR 2026 code and weights; modified ResNet-50; GAP feature of dimension 2048; 256 input; batch 64.	Frozen backbone followed by StandardScaler and logistic regression.
IncreFA	Paper-based adaptation rather than official code; OpenAI CLIP ViT-B/16; 256 input; latent dimension 512; batch 512; Adam, lr 10^{-3}; four epochs per task.	Final epoch; incremental replay with 150 samples per class; $\tau = 0.65$.

D.3 Baseline Implementations

All methods use the same known–unknown source partitions and the same training, validation, and test manifests within each protocol. Trainable baselines receive only the known-source training samples. Unknown-source labels are not used for supervised optimization.

For open-set evaluation, each method uses its own attribution output and unknown score. For auxiliary NCD evaluation, we extract the final attribution representation of each method, apply ℓ_2 normalization, and run the same K-means configuration ($k = 4$, `n_init=20`, seed 42). This keeps the clustering procedure fixed while comparing the source-related structure retained by different representations.

Table A.4 summarizes the baseline configurations. Adapted implementations use released code or weights; IncreFA is implemented from its published description.

D.4 Random Seeds, Hardware, and Runtime

All runs use seed 42 for Python, NumPy, PyTorch, CUDA, and K-means, with `cudnn.deterministic=true` and `cudnn.benchmark=false`.

The frozen intermediate-feature projector is initialized with Xavier uniform weights and is recreated in the training, prototype-construction, and evaluation processes. It is not stored in the model checkpoint. Exact reproduction therefore requires the same random seed and object initialization order.

The experiments use one NVIDIA H100 GPU with 80 GB of memory. Peak memory use is 26.2 GB, and mean training time is approximately 4,984 seconds (1.38 hours) per protocol.

E Complete Protocol-Wise Results

E.1 Main Open-Set Attribution Results

Tables A.6–A.15 report the complete protocol-wise results. Open-set attribution and unknown-source clustering are shown in separate single-column tables. The best and second-best values within each protocol are shown in **bold** and underlined, respectively.

Figure A.1 provides a qualitative view of the unknown-source compositional features used by the auxiliary NCD evaluation. The ground-truth source labels are applied only after representation learning and are used only to color the visualization.

Table A.5: Open-set attribution performance on GenImage.

Method	Closed-Set	Open-Set	
	ACC	AUC	OSCR
ResNet-50	88.24	65.17	63.82
DCT-CNN	79.53	63.49	62.05
RepMix	91.35	73.59	71.43
POSE	86.08	71.25	70.16
Abady et al.	79.91	67.33	65.47
Cioni et al.	91.67	77.82	75.39
FSD	92.12	83.85	83.46
RdH	**96.32**	**90.65**	**87.67**

Table A.6: Open-set attribution results on Balanced. Values are percentages.

Method	ACC	AUROC	FPR@95	OSCR
POSE	83.88	80.15	80.36	71.49
ASV	63.19	46.00	91.50	33.04
RepMix	93.87	84.04	70.94	81.46
Cioni et al.	98.81	93.45	37.85	**93.08**
FSD	91.20	89.37	33.75	65.11
LIDA	**99.95**	83.79	55.16	82.34
IncreFA	99.77	85.77	55.73	83.98
CLIP-Lin.	94.20	63.00	77.37	57.65
GNN	99.71	91.71	43.81	73.85
RdH	99.89	**95.70**	**20.21**	85.97

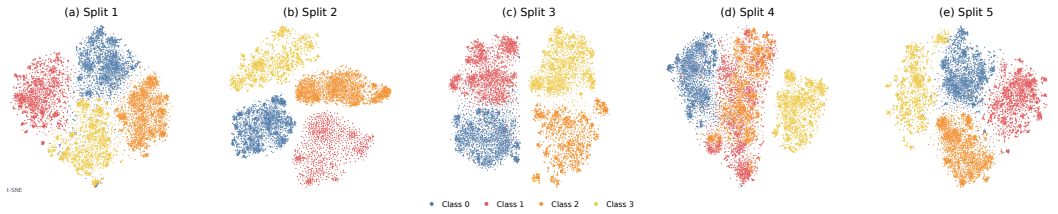


Figure A.1: t-SNE of unknown-source compositional features across the five protocols. Colors denote ground-truth source labels used only for post-hoc visualization.

GenImage results. Table A.5 reports the additional GenImage comparison referenced in the main paper. GenImage uses $K_H = 16$ for per-image restricted incidence and $K_p = 32$ for source prototypes in all five GenImage splits. AUC denotes the known-versus-unknown AUROC.

Across the five protocols, RdH obtains macro-average results of 98.64% ACC, 92.84% AUROC, 28.43% FPR@95, 84.50% OSCR, 79.15% NMI, and 78.90% ARI. Its AUROC varies from 90.58% to 95.70%, corresponding to a range of 5.12 percentage points.

E.2 Controlled Comparisons and Component Analysis

The protocol-wise results in Section E.1 include CLIP-Lin. and GNN comparisons under the same frozen CLIP representation.

The main paper reports the macro-average component analysis. Restricted incidence primarily improves open-set separation, while degree-aware weighting is particularly important for retaining unknown-source structure.

The complete protocol-wise component results are shown in Tables A.16–A.18. Protocol abbreviations are Bal. (Balanced), Diff. (Diff.-Known), Het. (Heterogeneous Mix), Lin. (Lineage), and U-D (U-Net→DiT).

Table A.7: Unknown-source clustering results on Balanced. Values are percentages.

Method	NMI	ARI
POSE	2.16	2.00
ASV	5.07	4.73
RepMix	1.92	1.67
Cioni et al.	34.32	31.35
FSD	7.63	6.09
LIDA	6.81	5.39
IncreFA	4.34	3.20
CLIP-Lin.	9.60	8.73
GNN	9.77	7.35
RdH	**69.86**	**75.38**

Table A.8: Open-set attribution results on Diff.-Known. Values are percentages.

Method	ACC	AUROC	FPR@95	OSCR
POSE	74.17	72.63	88.42	59.62
ASV	58.76	47.22	91.49	34.21
RepMix	74.32	54.01	85.18	48.20
Cioni et al.	79.76	65.58	87.89	59.98
FSD	78.78	70.21	48.32	68.91
LIDA	83.95	78.90	45.90	73.84
IncreFA	77.75	85.47	47.16	75.49
CLIP-Lin.	90.97	66.40	83.55	60.90
GNN	**99.79**	90.93	38.16	73.10
RdH	99.75	**94.35**	**20.71**	**86.36**

Table A.9: Unknown-source clustering results on Diff.-Known. Values are percentages.

Method	NMI	ARI
POSE	4.84	4.05
ASV	19.92	16.96
RepMix	4.61	3.23
Cioni et al.	89.34	86.17
FSD	23.27	17.36
LIDA	44.62	38.56
IncreFA	38.71	31.65
CLIP-Lin.	47.04	43.06
GNN	85.87	83.89
RdH	**97.32**	**98.00**

Table A.10: Open-set attribution results on Hetero. Mix. Values are percentages.

Method	ACC	AUROC	FPR@95	OSCR
POSE	70.81	71.99	89.42	57.43
ASV	52.15	44.14	94.53	25.33
RepMix	94.99	73.99	83.94	72.79
Cioni et al.	98.88	77.17	63.29	76.96
FSD	76.53	85.46	40.25	30.45
LIDA	91.93	87.25	44.44	84.57
IncreFA	97.93	83.37	52.62	80.29
CLIP-Lin.	88.21	68.86	75.53	62.95
GNN	99.31	72.68	87.20	67.94
RdH	**99.48**	**91.83**	**29.35**	**85.13**

Table A.11: Unknown-source clustering results on Hetero. Mix. Values are percentages.

Method	NMI	ARI
POSE	3.36	2.84
ASV	3.62	3.70
RepMix	24.78	23.47
Cioni et al.	79.01	79.13
FSD	5.48	4.32
LIDA	10.55	8.88
IncreFA	38.57	31.92
CLIP-Lin.	21.96	19.44
GNN	60.09	50.83
RdH	**83.83**	**85.15**

Table A.12: Open-set attribution results on Lineage. Values are percentages.

Method	ACC	AUROC	FPR@95	OSCR
POSE	56.13	61.44	92.95	41.03
ASV	52.12	55.55	91.29	33.98
RepMix	83.05	59.65	89.53	54.35
Cioni et al.	92.57	62.23	93.22	61.10
FSD	84.19	86.77	42.67	74.92
LIDA	75.51	67.39	91.98	62.11
IncreFA	87.85	76.24	95.15	70.93
CLIP-Lin.	78.95	68.44	69.53	60.69
GNN	**96.53**	81.31	54.59	67.90
RdH	94.32	**90.58**	**34.79**	**83.81**

Table A.13: Unknown-source clustering results on Lineage. Values are percentages.

Method	NMI	ARI
POSE	1.29	1.23
ASV	2.88	2.47
RepMix	26.77	21.39
Cioni et al.	62.72	**57.28**
FSD	17.35	21.39
LIDA	6.83	6.54
IncreFA	31.33	29.31
CLIP-Lin.	0.18	0.13
GNN	57.05	44.27
RdH	**69.45**	57.11

Table A.14: Open-set attribution results on U-Net→DiT. Values are percentages.

Method	ACC	AUROC	FPR@95	OSCR
POSE	74.17	70.55	89.64	58.27
ASV	58.92	46.47	91.05	34.24
RepMix	91.63	81.16	78.18	77.85
Cioni et al.	98.73	88.81	55.98	**88.47**
FSD	83.57	82.94	59.28	62.81
LIDA	83.95	77.77	57.50	72.98
IncreFA	93.88	73.08	54.26	69.12
CLIP-Lin.	90.92	50.61	88.63	49.15
GNN	**99.84**	**96.81**	**14.01**	74.74
RdH	99.75	91.76	37.08	81.24

Table A.15: Unknown-source clustering results on U-Net→DiT. Values are percentages.

Method	NMI	ARI
POSE	2.15	2.22
ASV	1.83	0.87
RepMix	6.20	6.33
Cioni et al.	<u>39.26</u>	<u>35.64</u>
FSD	15.61	12.37
LIDA	3.30	2.34
IncreFA	10.83	8.79
CLIP-Lin.	11.38	12.18
GNN	27.93	19.44
RdH	**75.27**	**78.84**

Table A.16: Protocol-wise component analysis: ACC and AUROC. Values are percentages.

Prot.	Variant	ACC	AUROC
Bal.	**RdH**	99.89	95.70
Bal.	Unres. incidence	99.84	95.99
Bal.	Uniform weighting	99.85	92.52
Bal.	No struct. loss	99.93	97.18
Bal.	Static atoms	99.85	92.74
Diff.	**RdH**	99.75	94.35
Diff.	Unres. incidence	99.76	90.50
Diff.	Uniform weighting	99.83	87.72
Diff.	No struct. loss	99.83	92.65
Diff.	Static atoms	99.79	91.04
Het.	**RdH**	99.48	91.83
Het.	Unres. incidence	99.53	91.10
Het.	Uniform weighting	99.39	80.15
Het.	No struct. loss	99.49	91.21
Het.	Static atoms	99.53	89.69
Lin.	**RdH**	94.32	90.58
Lin.	Unres. incidence	94.91	59.92
Lin.	Uniform weighting	94.84	85.67
Lin.	No struct. loss	95.39	71.39
Lin.	Static atoms	94.77	72.95
U-D	**RdH**	99.75	91.76
U-D	Unres. incidence	99.75	90.11
U-D	Uniform weighting	99.84	84.71
U-D	No struct. loss	99.83	94.52
U-D	Static atoms	99.76	85.12

Table A.17: Protocol-wise component analysis: FPR@95 and OSCR. Values are percentages.

Prot.	Variant	FPR@95	OSCR
Bal.	**RdH**	20.21	85.97
Bal.	Unres. incidence	17.97	86.02
Bal.	Uniform weighting	38.11	83.83
Bal.	No struct. loss	12.65	88.45
Bal.	Static atoms	35.39	83.19
Diff.	**RdH**	20.71	86.36
Diff.	Unres. incidence	38.45	81.51
Diff.	Uniform weighting	43.77	78.90
Diff.	No struct. loss	29.36	83.02
Diff.	Static atoms	31.79	83.34
Het.	**RdH**	29.35	85.13
Het.	Unres. incidence	28.04	83.88
Het.	Uniform weighting	74.77	72.04
Het.	No struct. loss	31.73	83.27
Het.	Static atoms	39.88	82.67
Lin.	**RdH**	34.79	83.81
Lin.	Unres. incidence	82.52	57.24
Lin.	Uniform weighting	68.89	79.08
Lin.	No struct. loss	64.25	66.41
Lin.	Static atoms	65.43	68.90
U-D	**RdH**	37.08	81.24
U-D	Unres. incidence	43.35	80.05
U-D	Uniform weighting	60.29	74.35
U-D	No struct. loss	24.91	84.82
U-D	Static atoms	58.85	76.01

Table A.18: Protocol-wise component analysis: NMI and ARI. Values are percentages.

Prot.	Variant	NMI	ARI
Bal.	**RdH**	69.86	75.38
Bal.	Unres. incidence	57.73	60.73
Bal.	Uniform weighting	49.46	48.54
Bal.	No struct. loss	64.97	70.25
Bal.	Static atoms	63.35	68.40
Diff.	**RdH**	97.32	98.00
Diff.	Unres. incidence	97.72	98.55
Diff.	Uniform weighting	87.31	86.19
Diff.	No struct. loss	97.71	98.36
Diff.	Static atoms	97.89	98.52
Het.	**RdH**	83.83	85.15
Het.	Unres. incidence	81.92	82.74
Het.	Uniform weighting	59.77	46.49
Het.	No struct. loss	82.03	83.05
Het.	Static atoms	81.12	81.90
Lin.	**RdH**	69.45	57.11
Lin.	Unres. incidence	67.20	63.66
Lin.	Uniform weighting	50.53	23.46
Lin.	No struct. loss	66.67	63.15
Lin.	Static atoms	61.99	56.78
U-D	**RdH**	75.27	78.84
U-D	Unres. incidence	66.89	71.53
U-D	Uniform weighting	49.51	51.29
U-D	No struct. loss	73.57	78.15
U-D	Static atoms	73.60	78.30

E.3 Alternative Visual Backbones

The backbone comparison uses the same feature-alignment module, invalid-image handling, and training schedule for all four encoders. The pretrained encoder is frozen, while a bias-free linear `FeatureInputAdapter` maps the extracted representation to 768 dimensions. The adapter, atom dictionary, hypergraph network, and classifier are retrained from scratch for 10 epochs with seed 42, and the checkpoint with the lowest validation loss is selected. The SigLIP backbone uses `vit_large_patch16_siglip_256.webli`; the local name `siglip2_vit_l16_256` is a directory alias rather than the model version. The exact non-CLIP weight identifiers are `vit_large_patch14_dinov2.lvd142m`, `vit_large_patch16_siglip_256.webli`, and `resnet50.tv2_in1k`.

Table A.19: Feature extraction and alignment for the alternative visual backbones.

Backbone	Model / weights	Extracted representation	Original dim.	Aligned dim.
CLIP	OpenAI ViT-L/14	Block-12 CLS token; final LN/projected CLS token	1024 / 768	768
DINOv2	ViT-L/14, LVD-142M	Block-12 CLS token; final normalized CLS token	1024	768
SigLIP	ViT-L/16-256, WebLI	Block-12 patch mean; final MAP-pooled feature	1024	768
ResNet-50	TV2-IN1K	Global-average-pooled layer-3 and layer-4 features	1024 / 2048	768

F Additional Analysis of Compositional Evidence

This section provides additional evidence that source information is carried by structured atom combinations rather than by a single source-exclusive atom. We report protocol-wise activation overlap, activation-ranked and mass-matched random deletion, single- and multi-atom sensitivity, and reuse-stratified equal-budget masking. Unknown-source labels are used only for post-hoc grouping and evaluation, except that the explicitly labeled "Worst-single" statistic is a post-hoc test-set sensitivity analysis.

F.1 Atom Combinations across Protocols

For each source, we average normalized incidence features over its evaluation images and select the ten atoms with the largest mean activations as its source signature.

Figure A.2 visualizes pairwise Jaccard overlap between the source-level Top-10 atom signatures. The matrices show substantial reuse both within and across the known and unknown groups. Their non-identical overlap patterns also indicate that the sources do not collapse to one common atom set.

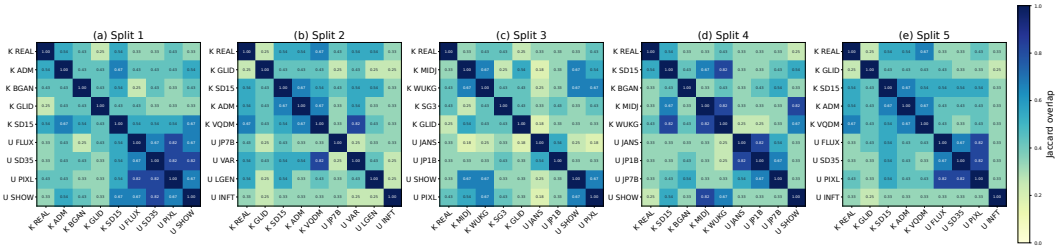


Figure A.2: Pairwise Jaccard overlap between source-level Top-10 atom signatures in the five protocols. K and U denote known and unknown sources, respectively.

Table A.20 reports, for every unknown source, how many of its Top-10 atoms also occur in the union of the known-source Top-10 signatures from the same protocol. Most unknown sources reuse seven to ten atoms already active for known sources. The unknown sources nevertheless differ in which atoms are combined and in their relative activation strengths. Thus, high atom reuse does not imply identical source fingerprints.

Table A.20: Overlap between each unknown-source Top-10 signature and the union of known-source Top-10 atoms in the same protocol. "Union" is the number of distinct atoms in the known-source union.

Protocol	Unknown Source	Overlap	Union
Balanced	FLUX	8/10	19
Balanced	PixArt-XL	8/10	19
Balanced	SD35-L	8/10	19
Balanced	Show-o	8/10	19
Diff.-Known	Infinity	7/10	19
Diff.-Known	Janus-Pro-7B	8/10	19
Diff.-Known	LlamaGen	10/10	19
Diff.-Known	VAR	10/10	19
Hetero. Mix	Janus	7/10	19
Hetero. Mix	Janus-Pro-1B	9/10	19
Hetero. Mix	PixArt-XL	9/10	19
Hetero. Mix	Show-o	9/10	19
Lineage	Janus	7/10	17
Lineage	Janus-Pro-1B	8/10	17
Lineage	Janus-Pro-7B	9/10	17
Lineage	Show-o	9/10	17
U-Net→DiT	FLUX	9/10	19
U-Net→DiT	Infinity	7/10	19
U-Net→DiT	PixArt-XL	9/10	19
U-Net→DiT	SD35-L	9/10	19

F.2 Activation-Ranked and Mass-Matched Random Deletion

For each Cross-Era protocol, atoms are ranked globally using the class-balanced mean of $\mathbf{H}_{\mathrm{norm}}$ over the known-source training set, including the real-image class. The ranking is computed independently for each protocol and does not use test labels. Ranked deletion masks the Top-n atoms for $n \in \{1, 2, 4, 8\}$ by setting the corresponding columns of $\mathbf{H}_{\mathrm{raw}}$ to zero. We then recompute $\mathbf{H}_{\mathrm{norm}}$, hypergraph propagation, classification logits, and the unknown score.

The frozen-prototype condition retains the prototypes built from the unmodified training set. The rebuilt-prototype condition applies the same mask to the training representations, reselects the Top-K_p surviving dimensions, and normalizes the reconstructed prototypes. The random control matches the total training activation mass removed by the ranked intervention rather than the number of deleted atoms. Each random budget uses 20 independently sampled masks with base seed 271828. Variation across these masks is not a bootstrap confidence interval over evaluation images.

Table A.21: Five-protocol macro-average prototype AUROC under the Top-8 ranked deletion and activation-mass-matched random control. Drops are percentage points relative to the shared prototype baseline.

Deletion	Prototype	Baseline	After deletion	Drop
Ranked Top-8	Frozen	84.717	71.894	**12.823**
Ranked Top-8	Rebuilt	84.717	71.365	**13.352**
Random mass-matched	Frozen	84.717	78.544	**6.173**
Random mass-matched	Rebuilt	84.717	79.496	**5.221**

Because matching is based on activation mass, the random control is not an equal-cardinality intervention. At the Top-8 ranked budget, a random mask removes 50.92 atoms on average, with protocol-level averages ranging from approximately 31 to 68 atoms. This distinction is necessary when interpreting the larger degradation produced by the eight highly activated atoms.

F.3 Single- and Multi-Atom Deletion

The Top-n conditions use the same known-training class-balanced ranking and rebuilt-prototype procedure described above. They are global protocol-level masks, not per-image or unknown-source-conditioned masks. "Worst-single" instead deletes each of the 128 atoms in turn and retrospectively reports, for each protocol, the atom producing the largest test prototype-AUROC decrease.

It is therefore a post-hoc test-set sensitivity statistic rather than a validation-selected intervention or a deployable selection rule.

Table A.22: Prototype AUROC under single- and multi-atom deletion with rebuilt prototypes. "Worst-single" is selected post hoc on each protocol's test set.

Protocol	Baseline	Worst-single	Top-1	Top-2	Top-4	Top-8
1	85.147	83.153	83.313	82.898	82.087	72.786
2	82.318	78.331	78.331	78.984	75.249	72.392
3	83.549	81.351	81.636	81.152	76.284	65.959
4	87.430	85.162	85.162	85.002	82.561	70.001
5	85.140	82.912	84.851	85.432	83.900	75.688
Mean	84.717	82.182	82.659	82.694	80.016	71.365
Mean drop	–	**2.535**	**2.058**	**2.023**	**4.701**	**13.352**

The mean worst-single prototype-AUROC decrease is $(1.994+3.987+2.198+2.268+2.228)/5 = 2.535$ percentage points.

The largest single-atom prototype-AUROC decrease is below five percentage points in each protocol.

F.4 Reuse-Stratified Atom Masking

We further test whether atoms with different reuse levels contribute equally to source evidence. Reuse groups are defined using the known-source training data:

$$\begin{aligned} \text{Private}: &\quad r_j = 1, \\ \text{Moderate}: &\quad r_j = 2, \\ \text{Broad}: &\quad r_j \geq 3, \end{aligned} \tag{A.91}$$

where r_j is the number of known sources whose source-level support contains atom j.

For each group, we mask four atoms from the evaluation representations, renormalize the remaining incidence feature, rebuild the known-source prototypes, and recompute the metrics. Sampling is balanced across source memberships and repeated for 20 seeds.

The test-set prototype-AUROC changes plotted in the main paper are $+0.64$, -2.80, and -1.33 percentage points for private, moderate, and broad reuse, respectively.

G Visual Summaries of Compositional Evidence

G.1 Atom-Associated Visual and Spectral Statistics

Figures A.3–A.7 show selected atom-induced hyperedges for each protocol. Each column contains high-activation images, an edge-specific frequency map, the mean Fourier spectrum, a coarse high-frequency grid, and frequency-band shares. Hyperedges are selected using normalized incidence statistics without unknown-source labels.

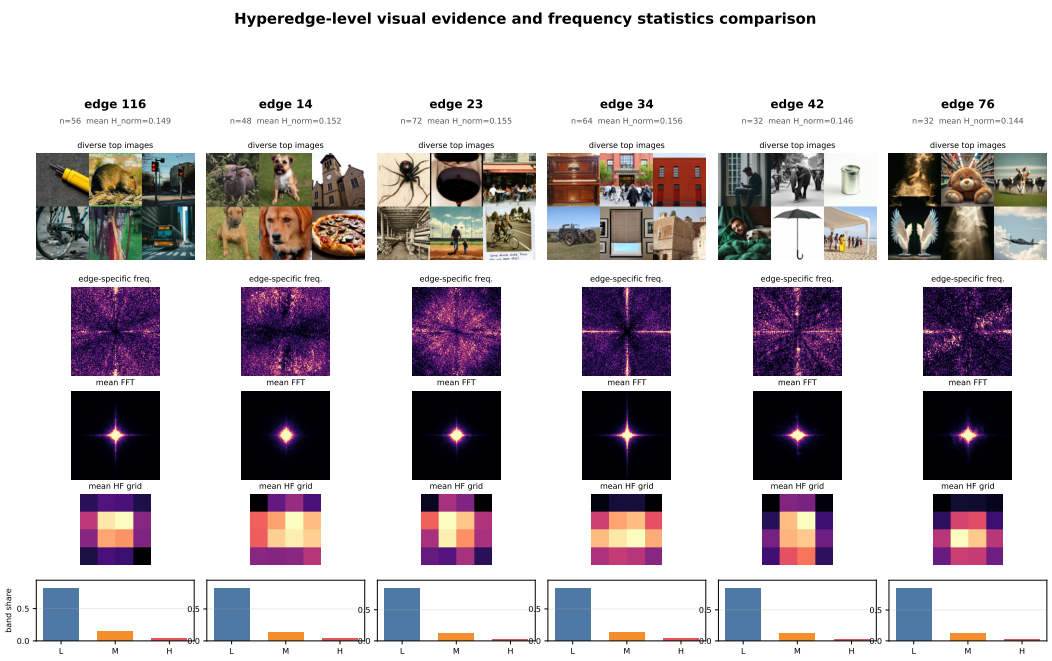


Figure A.3: Hyperedge-level visual and spectral summaries for S1 (Balanced).

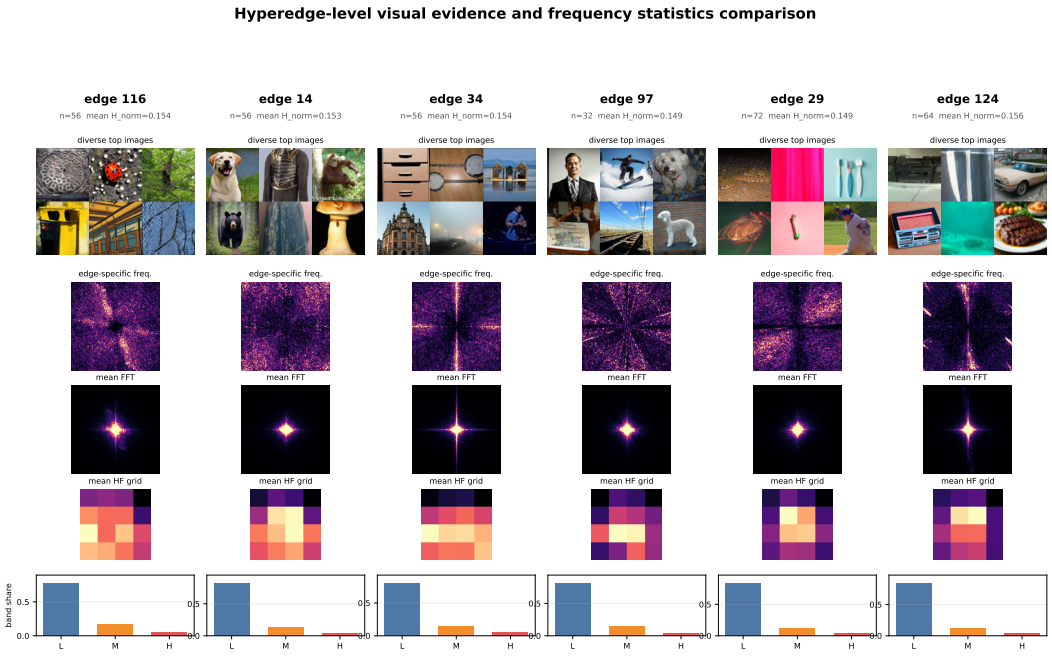


Figure A.4: Hyperedge-level visual and spectral summaries for S2 (Diff.-Known).

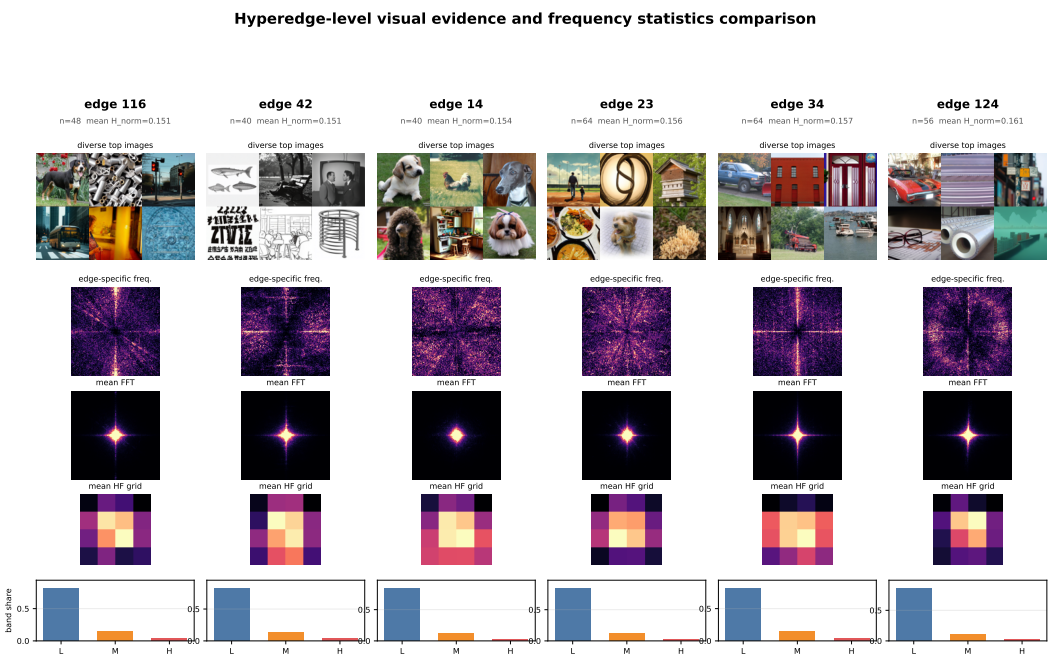


Figure A.5: Hyperedge-level visual and spectral summaries for S3 (Hetero. Mix).

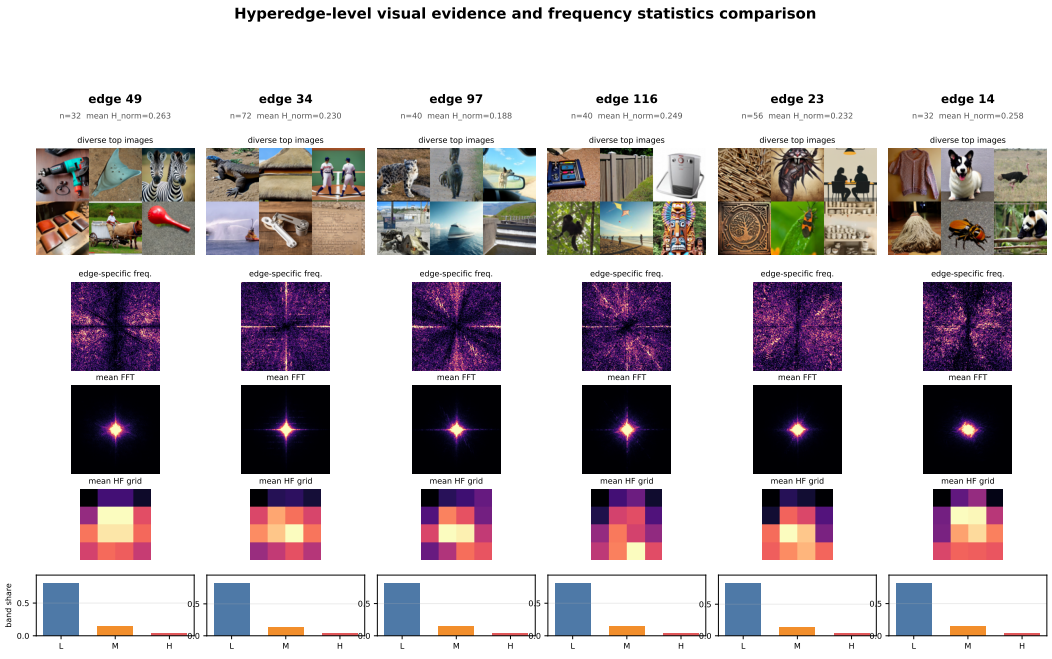


Figure A.6: Hyperedge-level visual and spectral summaries for S4 (Lineage).

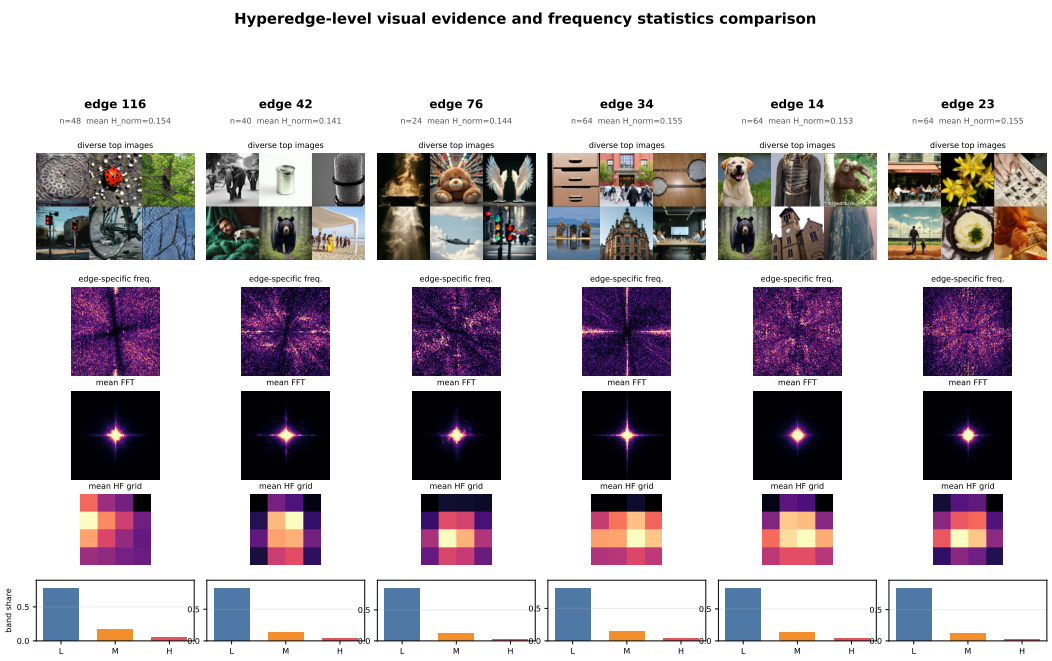


Figure A.7: Hyperedge-level visual and spectral summaries for S5 (U-Net→DiT).

H Propagation, Batch, and Input Stability

This section examines three practical properties of RdH: its behavior under repeated propagation, its sensitivity to test-batch composition, and its robustness to common input perturbations.

H.1 Propagation-Depth Analysis

The main paper evaluates propagation depth. Depth 0 denotes the input representation before propagation; RdH uses six layers in the main experiments.

H.2 Test-Batch Sensitivity

The incidence feature of an image is computed independently under the fixed atom dictionary, but its propagated classification feature depends on the other images in the test batch. We therefore evaluate the threshold-free stability of candidate attribution and unknown scores under repeated batch regrouping. The model remains in evaluation mode, and all learned parameters and source prototypes remain fixed. Consequently, the analysis isolates the effect of batch size and composition without introducing threshold-selection uncertainty.

For each Cross-Era protocol, 180 anchor images are evaluated under batch sizes

$$B \in \{8, 16, 32, 64, 128\} \tag{A.92}$$

and known-to-unknown compositions

$$3{:}1, \qquad 1{:}1, \qquad 1{:}3. \tag{A.93}$$

Each anchor is combined with independently sampled companion images in ten repeated regroupings for every condition. We report raw attribution agreement and the mean per-anchor standard deviation of the unknown score. Raw attribution agreement measures pairwise consistency of the candidate known-source prediction across regroupings. Both quantities are independent of a binary known–unknown operating threshold.

Table A.23: Threshold-free test-batch stability as a function of batch size, averaged over five protocols and three known-to-unknown compositions. Raw attribution agreement is reported as a percentage, and score standard deviation is multiplied by 10^3.

B	Raw attribution agreement	$10^3 \times$Score std
8	95.33	1.359
16	97.07	0.984
32	99.03	0.327
64	99.72	0.085
128	99.86	0.016

The threshold-free quantities become more stable as the batch size increases. Raw attribution agreement rises from 95.33% at $B = 8$ to 99.86% at $B = 128$. Over the same range, the mean score standard deviation decreases from 1.359×10^{-3} to 1.6×10^{-5}. Therefore, the batch effect is concentrated primarily in small-batch candidate attribution.

H.3 Input Corruption Settings and Results

We evaluate common post-processing perturbations on the Balanced protocol. For each condition, 50% of the test images are selected with a fixed random seed and perturbed independently; the remaining images are unchanged. All methods use the same corrupted test manifest.

The perturbation settings are:

- **Brightness, contrast, and saturation:** multiplicative factor sampled from $[0.85, 1.15]$.
- **Gaussian noise:** zero-mean noise with standard deviation 15 on the 0–255 pixel scale.
- **Gaussian blur:** blur radius 2.0.
- **Rotation:** angle sampled from $[-10°, 10°]$.
- **JPEG:** encoding quality 60.

Tables A.24 and A.25 provide the numerical results. The best and second-best values in each row are shown in **bold** and underlined, respectively.

Table A.24: AUROC under common input corruptions on Balanced. Values are percentages.

Corruption	RdH	Cioni	LIDA	IncreFA	POSE
Brightness	**81.97**	75.74	49.05	76.08	53.86
Contrast	**79.82**	76.25	41.68	71.53	53.83
Saturation	**80.23**	75.88	38.02	70.53	53.89
Noise	71.74	**78.16**	15.69	78.09	53.78
Blur	**82.31**	80.65	44.08	79.67	53.90
Rotation	**76.27**	70.60	36.95	71.39	43.35
JPEG	**90.79**	89.01	43.83	90.01	53.84

Table A.25: Unknown-source NMI under common input corruptions on Balanced. Values are percentages.

Corruption	RdH	Cioni	LIDA	IncreFA	POSE
Brightness	**62.55**	59.03	30.58	18.29	4.69
Contrast	**67.84**	61.28	33.52	13.25	6.62
Saturation	**68.41**	62.43	31.23	13.34	6.25
Noise	**58.22**	53.41	43.61	21.58	4.27
Blur	**64.98**	61.04	39.85	23.16	7.81
Rotation	**71.05**	60.23	32.58	21.31	3.13
JPEG	**79.02**	69.02	23.54	23.39	4.69

RdH obtains the highest AUROC in six of seven tested corruption conditions and the highest NMI in all seven. Under Gaussian noise, Cioni and IncreFA obtain higher AUROC.