CROSS-LAYER RELATIONAL REPRESENTATION MODELING FOR FEW-SHOT AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated images have become increasingly realistic, raising urgent concerns about content authenticity and misuse, making reliable detection under unseen generators increasingly important. Since collecting large-scale training data for every emerging generator is impractical, this challenge motivates few-shot AI-generated image (AIGI) detection, where a detector must adapt to novel generators from only a handful of labeled examples. In this work, we identify generator-dependent cross-layer relational patterns in pretrained CLIP representations as a generalizable cue for AIGI detection. Specifically, real images and images from different generators exhibit distinct cross-layer behaviors in token representations and class-patch relational patterns. Motivated by this analysis, we propose a Cross-Layer Relational Module (CLRM), which constructs relational representations through bidirectional cross-attention between class tokens and patch tokens from intermediate and final layers. CLRM then transforms such semantic-spatial correspondence patterns across layers into metric-space embeddings and combines them with prototypical learning for few-shot adaptation. Extensive experiments demonstrate that CLRM achieves state-of-the-art performance in few-shot classification and generator source attribution, while providing strong generalization and robustness.

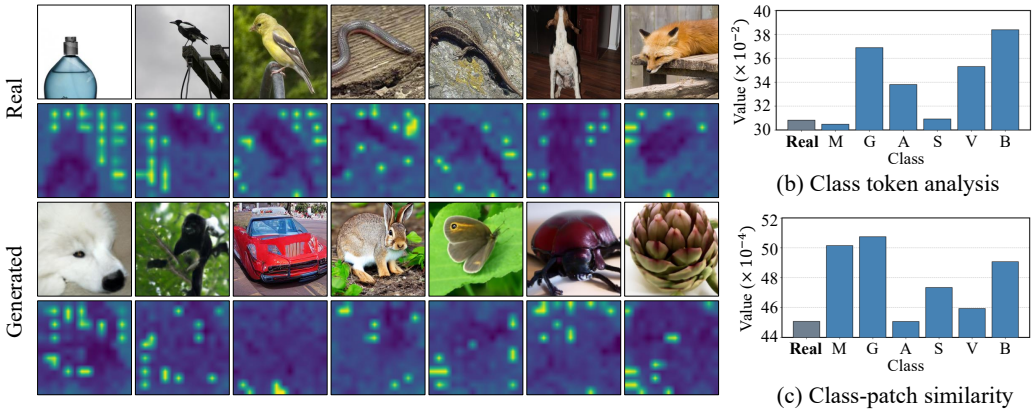
1 INTRODUCTION

Recent advances in generative models have gone far beyond producing photorealistic images that are visually indistinguishable from human-made content, emerging as major disruptors across industries that have traditionally relied on human creativity (Karras et al., 2020; Ho et al., 2020; Rombach et al., 2022; Saharia et al., 2022; Sauer et al., 2023; Binyamin et al., 2025). These are increasingly integrated into creative workflows, ranging from automated scene synthesis and cinematic style transfer in film production (Zhao & Zhao, 2024) to the large-scale generation of marketing visuals with remarkable speed, cost-efficiency, and even superhuman effectiveness (Hartmann et al., 2025).

However, the proliferation of synthetic content raises significant concerns regarding authenticity, misinformation, and potential misuse. These risks are no longer hypothetical: recent real-world incidents (Associated Press, 2024; CBS News, 2025; Yahoo Finance, 2025) have vividly illustrated the societal impact of maliciously deployed generative models. Consequently, the development of reliable detectors for AI-generated content has become an urgent necessity (Zhang et al., 2019; Tan et al., 2023; Chen et al., 2024; Ricker et al., 2024; Jia et al., 2025; Wang et al., 2026).

Despite substantial progress, current AI-generated image (AIGI) detection approaches still struggle to generalize across diverse generative models. Detectors trained on a fixed set of generators often fail when confronted with images produced by unseen generators, as each generator induces its own distribution shaped by its unique architecture and objective (Cazenavette et al., 2024). A naïve strategy of continually collecting new data and retraining detectors for every emerging generator is infeasible as generative models evolve rapidly and are increasingly commercialized, limiting access to large-scale samples for training. These constraints motivate a more realistic setting in which a detector must adapt to a novel generator using only a handful of labeled examples, i.e., few-shot AIGI detection (Wu et al., 2025).

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.



(a) Feature differences between the final and intermediate layer features

(b) Class token analysis

(c) Class-patch similarity transition analysis

Figure 1: Cross-layer behaviors between real and generated images. Each label (M, G, A, S, V, B) corresponds to a generator, i.e., initial letter of each generator. (a) Feature differences between the final and intermediate CLIP layers, showing different spatial transition patterns for real and generated images. (b) Similarity analysis of the class token between final and intermediate layers, revealing generator-dependent transition. (c) Class-patch similarity transition analysis. We compute the patch-wise differences between final and intermediate layer class-patch similarities and measure their variance across patches, showing that cross-layer class-patch correspondence shifts differently across generators.

A key question then arises: what representation enables few-shot generalization across generators? Pretrained vision-language models such as CLIP (Radford et al., 2021) offer strong transferable features and have been shown to generalize effectively for AI-generated image detection (Ojha et al., 2023). However, using CLIP as a single final layer embedding can obscure how local visual evidence is transformed into a global semantic summary. AI-generated images may appear globally plausible while retaining local, low-level irregularities that can be attenuated in final layer semantic abstraction (Chai et al., 2020; Corvi et al., 2023; Koutlis & Papadopoulos, 2024). This makes the layer-wise behavior of CLIP particularly relevant for AIGI detection: its image representation evolves across model layers, and intermediate representations can provide information complementary to final-layer features (Gandelsman et al., 2024; Chen et al., 2025). Our empirical analysis in Fig. 1 further reveals that real images and images from different generative models exhibit distinct layer-wise transition behaviors in patch tokens, class tokens, and class-patch similarities. These behaviors suggest that each generator induces a characteristic cross-layer relational pattern between semantic summaries and spatial evidence. Exploiting these subtle semantic-spatial correspondence patterns can yield a more robust representation for few-shot AIGI detection than relying solely on a single-layer embedding.

Based on this analysis, we propose a Cross-Layer Relational Module (CLRM) that explicitly models cross-layer relations within CLIP for few-shot AIGI detection. The class-token transition itself is generator-dependent, while patch-level discrepancy maps and class-patch similarity shifts reveal spatially varying relational patterns. Therefore, CLRM constructs relational representations by using class tokens as semantic queries and patch tokens as spatial evidence across layers. Specifically, given a few-shot episode, we extract the intermediate and final CLIP features for both support and query images, each consisting of a class token and patch tokens. CLRM then performs bidirectional cross-attention between the class token of one layer and the patch tokens of the other layer to explicitly encode cross-layer relations, and fuses the resulting relational features into a unified embedding. This embedding is incorporated into a prototypical metric learning framework, where class prototypes are computed from the support set and query samples are classified by distances to these prototypes. By transforming generator-dependent cross-layer relational patterns into metric-space embeddings, our detector can adapt to unseen generators with limited labeled examples. Extensive experiments and analyses demonstrate that our method not only achieves state-of-the-art accuracy but also exhibits strong generalization.

In summary, our contributions are: (i) We reveal that real images and images from different generators exhibit generator-dependent cross-layer behaviors in CLIP, including feature transitions and class-patch relational shifts, providing a transferable cue for few-shot AIGI detection. (ii) We introduce the Cross-Layer Relational Module (CLRM), which explicitly models bidirectional class-to-

patch relations between intermediate and final CLIP layers and converts semantic-spatial correspondence patterns into discriminative metric-space embeddings. (iii) We demonstrate state-of-the-art performance on few-shot binary classification and generator source attribution, with additional robustness under dataset shifts, input perturbations, and noisy support labels.

2 Related Work

2.1 AI-Generated Image Detection

Early studies on AI-generated image detection targeted low-level artifacts introduced by generative models (Zhang et al., 2019; Wang et al., 2020; Qian et al., 2020; Liu et al., 2020). However, their reliance on generator-specific cues limited their generalization to unseen models. To improve generalization, DIRE (Wang et al., 2023) and LARE2 (Luo et al., 2024) exploit reconstruction characteristics to distinguish real and AI-generated images. Foundation-model-based detectors have recently shown strong promise for cross-generator generalization. UFD (Ojha et al., 2023) demonstrates that frozen CLIP features provide strong transferability across unseen generators. However, relying solely on the final layer CLIP embeddings can overlook intermediate visual cues that remain useful for identifying generated artifacts. RINE (Koutlis & Papadopoulos, 2024) addresses this limitation by leveraging intermediate class tokens with learnable importance weights, showing that intermediate CLIP representations contain valuable cues for AI-generated image detection. While these approaches highlight the importance of intermediate representations, they focus on selecting or aggregating informative layer-wise features. In contrast, our work explores cross-layer relations by considering how semantic and spatial representations evolve across layers through both class and patch tokens. This distinction is important for few-shot AIGI detection and generator source attribution, where fine-grained generator sources must be distinguished under limited supervision.

2.2 Few-Shot Learning

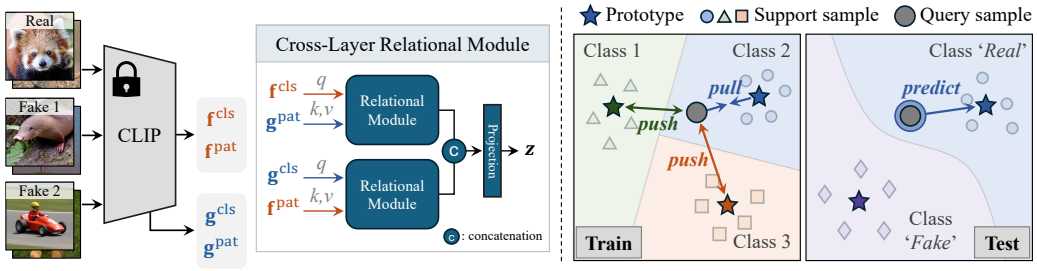
Few-shot learning (FSL) aims to classify novel categories using only a handful of labeled samples (Snell et al., 2017; Simon et al., 2020; Xie et al., 2022; Lee et al., 2022; Hao et al., 2023). This paradigm aligns naturally with AI-generated image detection, where new generators emerge frequently and collecting large-scale labeled data for each is infeasible (Wu et al., 2025). Metric-based FSL approaches, such as Prototypical Networks (Snell et al., 2017), learn transferable embedding spaces by comparing similarities between support and query instances instead of retraining full classifiers. Inspired by this principle, we design a few-shot detection framework that leverages the relational gap between intermediate and final CLIP representations. By explicitly encoding these cross-layer discrepancies, our method enhances the separation between real and AI-generated images, leading to improved detection performance and stronger generalization.

3 Method

3.1 Overview

The overall architecture of our method is illustrated in Fig. 2. Given an episode composed of support and query samples, we first extract both intermediate and final layer features from a Vision Transformer (ViT) (Dosovitskiy et al., 2021) pretrained with CLIP (Radford et al., 2021), where each representation consists of a class token and patch tokens. As illustrated in Fig. 1, the discrepancy between intermediate and final CLIP features provides a distinctive cue for distinguishing AI-generated images from real ones. We exploit this observation by introducing a Cross-Layer Relational Module (CLRM) that explicitly models such cross-layer discrepancies to form discriminative relational representations for few-shot AI-generated image (AIGI) detection.

The following subsections describe (i) few-shot classification formulation for AIGI detection, (ii) the proposed relational representation modeling, (iii) the few-shot metric learning framework built upon it, and (iv) discussion of design rationale underlying our framework.



(a) Cross-Layer Relational Representation Modeling (b) Few-Shot Metric Learning

Figure 2: Illustration of our framework. (a) The proposed Cross-Layer Relational Module (CLRM) models bidirectional cross-attention between class and patch tokens across layers, encoding these cross-layer relationships into class-discriminative representations. (b) During training, the model learns a metric space where samples from the same class are pulled closer toward their class prototype, while samples from different classes are pushed apart within each episode. At test time, a new episode is constructed using support samples from unseen generators (and real images), and each query sample is classified based on its distance to the prototypes in the learned embedding space.

3.2 Problem Formulation

Few-shot classification aims to generalize to novel classes unseen during training using only a few labeled examples. Unlike conventional AIGI detection approaches that aggregate all AI-generated images produced by different generators into a single fake category, our formulation, following FSD (Wu et al., 2025), treats images produced by different generative models as distinct classes. This setup encourages the detector to learn fine-grained generator-specific characteristics rather than relying solely on a binary real–fake boundary.

Formally, each training episode consists of N classes, including $N-1$ randomly sampled fake classes and one real class. Each class contains K labeled samples forming the support set $S = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N\times K}$ where $\mathbf{x}$ and $\mathbf{y}$ denote input image and its corresponding label, respectively, defining an N-way K-shot task. Given the support set, the objective is to predict the correct label for each query sample in the query set $Q = \{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^{N\times K}$, which is disjoint from S but drawn from the same N classes. During testing, the support set $S^{\text{test}} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{\tilde{N}\times\tilde{K}}$ provides limited supervision for adaptation, and the query images from the unseen generator are used to evaluate generalization performance. Here, $\tilde{N} = 2$ corresponds to the conventional AIGI detection setting (i.e., binary classification), and $\tilde{K}$ denotes the number of shots at test time. Unless otherwise specified, we set $\tilde{K} = K$ in all experiments, except in ablation studies where we vary $\tilde{K}$ to analyze the impact of support set size.

3.3 Relational Representation Modeling

The motivation behind our relational representation modeling framework arises from the layer-wise behavior of CLIP representations (Radford et al., 2021; Li et al., 2024; Gandelsman et al., 2024; Chen et al., 2025). CLIP is trained through image-text alignment, and its final image representation provides a semantic summary, while intermediate tokens can preserve spatial and structural cues that are complementary to the final embedding (Chen et al., 2025). Such intermediate cues are particularly relevant for AIGI detection, since AI-generated images may appear globally plausible while still containing local or low-level irregularities that can be weakened after final layer semantic abstraction (Chai et al., 2020; Corvi et al., 2023; Koutlis & Papadopoulos, 2024). Consistent with this view, Fig. 1 shows that real images and images from different generators exhibit distinct cross-layer behaviors, including generator-dependent class-token transition profiles, spatially structured patch-token differences, and heterogeneous class-patch similarity shifts. These observations motivate us to model the relation between global semantic summaries and local spatial evidence across layers, rather than relying on a single-layer embedding.

To explicitly capture this signal, we extract intermediate and final layer features of an image $\mathbf{x}$ from CLIP encoder $\mathcal{F}$:

$$(\mathbf{f}, \mathbf{g}) = ([\mathbf{f}^{\text{cls}}, \mathbf{f}^{\text{pat}}], [\mathbf{g}^{\text{cls}}, \mathbf{g}^{\text{pat}}]) = \mathcal{F}(\mathbf{x}), \tag{1}$$

where $\mathbf{f}, \mathbf{g} \in \mathbb{R}^{(1+P)\times D}$ denote the final layer and intermediate layer token embeddings, respectively, each composed of one class token ($\mathbf{f}^{\text{cls}}, \mathbf{g}^{\text{cls}} \in \mathbb{R}^D$) and P patch tokens ($\mathbf{f}^{\text{pat}}, \mathbf{g}^{\text{pat}} \in \mathbb{R}^{P\times D}$) with embedding dimension D.

Given that our analysis indicates discrepancies between global semantics and spatial evidence across layers, we adopt cross-attention as a principled mechanism to model such cross-layer relations. Specifically, the class token of one layer is treated as a global query attending to the patch tokens of the other layer, measuring how high-level semantic summaries align with spatial representations extracted at a different level of hierarchy. We formally define the cross-attention operation:

$$\mathrm{CA}(\mathbf{q}, \mathbf{k}, \mathbf{v}) = \mathrm{Softmax}\left(\frac{\mathbf{q}\mathbf{k}^\top}{\sqrt{D}}\right)\mathbf{v}, \tag{2}$$

where $\mathbf{q} \in \mathbb{R}^{1\times D}$ is the *query*, $\mathbf{k}, \mathbf{v} \in \mathbb{R}^{P\times D}$ are the *key* and *value* matrices. We perform this operation bidirectionally to produce cross-layer relational features $\bar{\mathbf{f}}, \bar{\mathbf{g}} \in \mathbb{R}^{1\times D}$:

$$\bar{\mathbf{f}} = \mathrm{CA}(\mathbf{f}^{\text{cls}}, \mathbf{g}^{\text{pat}}, \mathbf{g}^{\text{pat}}), \tag{3}$$

$$\bar{\mathbf{g}} = \mathrm{CA}(\mathbf{g}^{\text{cls}}, \mathbf{f}^{\text{pat}}, \mathbf{f}^{\text{pat}}). \tag{4}$$

The two directions capture complementary forms of semantic-visual correspondence. In Eq. 3, the final class token attends to intermediate patch tokens, measuring whether the final semantic summary is supported by spatial evidence retained at an earlier layer. In Eq. 4, the intermediate class token attends to final layer patch tokens, measuring how an earlier semantic summary relates to spatial evidence after deeper CLIP processing. Since CLIP representations at different layers encode different levels of semantic and spatial information (Gandelsman et al., 2024; Chen et al., 2025), these two relations are not necessarily redundant. The bidirectional formulation allows CLRM to condition intermediate and final semantic summaries on spatial evidence from the opposite layer, thereby capturing complementary cross-layer relational patterns. The relational representations are then concatenated and projected to obtain a unified representation:

$$\mathbf{z} = \mathrm{Proj}\big(\mathrm{Concat}[\bar{\mathbf{f}}, \bar{\mathbf{g}}]\big), \tag{5}$$

where $\mathrm{Proj}(\cdot)$ is a learnable linear transformation.

By transforming cross-layer discrepancies into relational embeddings, CLRM amplifies hierarchical inconsistencies that are subtle in single-layer features but become discriminative when modeled relationally. The resulting representation is used for both support and query samples and serves as the foundation for downstream few-shot metric learning.

3.4 Few-Shot Metric Learning

To construct a transferable embedding space that generalizes well to unseen generators, we adopt the Prototypical Networks (Snell et al., 2017) objective as our metric learning framework for few-shot AIGI detection. This paradigm learns a discriminative metric space in which samples from the same class cluster closely, while samples from different classes are well separated.

Given an N-way K-shot episode with the support set S, the prototype representation for each class $n \in \{1, \ldots, N\}$ is computed as the mean embedding of its K support samples:

$$\mathbf{c}_n = \frac{1}{K}\sum_{i:\mathbf{y}_i=n} \mathbf{z}_i, \tag{6}$$

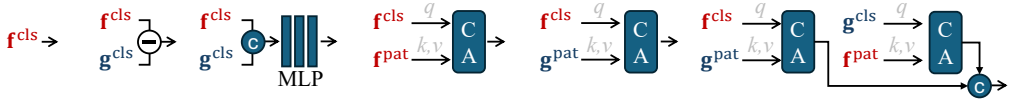
where $\mathbf{z}_i$ is the cross-layer relational representation extracted from our CLRM. Each prototype $\mathbf{c}_n$ thus serves as the representative feature of class n in the learned embedding space.

For each query sample $\mathbf{x}_q \in Q$, we compute its relational representation $\mathbf{z}_q$ and evaluate class probabilities via a distance-based softmax:

$$p_\theta(\mathbf{y} = n \mid \mathbf{x}_q) = \frac{\exp(-d(\mathbf{z}_q, \mathbf{c}_n))}{\sum_{m=1}^{N}\exp(-d(\mathbf{z}_q, \mathbf{c}_m))}, \tag{7}$$

where $d(\cdot, \cdot)$ denotes the Euclidean distance. The training objective minimizes the negative log-likelihood over all query samples:

$$\mathcal{L} = -\sum_{(\mathbf{x}_q, \mathbf{y}_q)\in Q} \log p_\theta(\mathbf{y}_q \mid \mathbf{x}_q). \tag{8}$$



(a) Final Layer (b) Direct Difference (c) Cross-Layer Implicit Modeling (d) Final-Layer Attention Modeling (e) Uni-directional Relation Modeling (f) Bi-directional Relation Modeling (CLRM, Ours)

Figure 3: Architectural variants to analyze CLRM. (a) Class token of the final layer. (b) Class token difference between the final layer and intermediate layer. (c) Implicit modeling by concatenating class tokens from the final and intermediate layers. (d) Final-layer attention using the class token as the *query* and the patch tokens as *keys* and *values*. (e) Uni-directional cross-layer attention where the *key* and *value* tokens are taken from the intermediate layer. (f) Our CLRM design.

This episodic training scheme encourages the model to learn a transferable relational embedding that generalizes to unseen generators with only a few examples. During inference, the same relational encoding and prototype computation are performed on the support set, and query samples are classified by proximity to the learned prototypes. By leveraging CLRM for representation extraction, our few-shot metric learning framework effectively captures cross-layer relation patterns of AI-generated images, yielding enhanced discriminability and robustness against unseen generative models.

3.5 Discussion

Design Rationale for CLRM. CLRM is designed as a semantic-visual consistency tester rather than a generic feature fusion module. The motivation in Fig. 1 indicates that the useful cue is not only the inter-layer difference, but also the relational transition between semantic summaries and spatial evidence. Accordingly, the architectural variants in Fig. 3 isolate the role of each design choice.

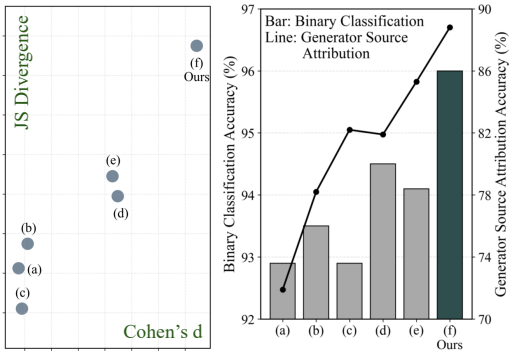


Figure 4: (Left): Separability between real and AI-generated images with Cohen's d and JS divergence, (Right): Few-shot performance of the architectural variants.

Variant (a) simply utilizes the final layer representation of CLIP (i.e., UFD (Ojha et al., 2023)). Variant (b) directly computes the difference between intermediate and final class tokens. This captures a coarse representation-level transition. (c) concatenates class tokens and models their relation with an MLP, testing whether implicit global aggregation is sufficient. However, it cannot explicitly model semantic-spatial correspondence. (d) introduces class-to-patch attention within the final layer, but this relation is computed only after CLIP has already formed a high-level representation, where subtle intermediate cues may be attenuated. (e) extends attention across layers by letting a class token attend to patch tokens from another layer, directly modeling cross-layer class-patch correspondence. Finally, our full CLRM (f) adopts bidirectional cross-layer attention. This design captures both directions of mismatch: whether intermediate semantic cues align with final spatial evidence, and whether final semantic summaries are supported by intermediate spatial evidence. Since the two directions condition different semantic summaries on spatial evidence from different abstraction levels, they are not necessarily redundant. The bidirectional formulation therefore provides a richer relational embedding than a single direction design.

Separability Analysis. To analyze these behaviors, Fig. 4 (left) visualizes the distributional separability between real and AI-generated images using Cohen's d (Cohen, 2013) and Jensen–Shannon divergence (Lin, 1991). Both metrics are averaged across six generators in GenImage (Zhu et al., 2023). We observe that as cross-layer modeling becomes more explicit from (a) to (f), separability consistently increases, with CLRM achieving the largest margin. This trend is also reflected in few-shot performance (Fig. 4, right) (detailed in the appendix). These results indicate that our CLRM not only improves end-to-end accuracy but also enlarges inter-class margins in the learned metric space.

Table 1: Performance on the GenImage dataset under the 3-way 5-shot setting. We adopt a leave-one-out setting for binary classification and generator source attribution, and leave-two-out for unseen generator source attribution. Binary classification distinguishes the excluded subset from real images, while generator source attribution classifies all generator sources and real images. M, G, A, S, V, and B denote Midjourney, GLIDE, ADM, Stable Diffusion, VQDM, and BigGAN, respectively.

(a) Binary classification.

Method	Excluding Subset						Avg.
	M	G	A	S	V	B	Acc.(%)
Spec	50.0	64.7	52.8	56.1	56.5	63.0	57.2
CNNSpot	52.8	73.3	55.0	55.9	54.4	66.2	59.6
F3Net	50.1	52.5	66.4	57.0	59.4	50.4	56.0
GramNet	51.3	62.6	53.8	56.8	52.2	57.2	55.7
DIRE	57.9	68.2	57.3	58.2	59.6	50.8	58.7
LARE2	62.7	80.2	63.5	79.6	76.9	72.0	72.5
UFD	86.7	98.6	90.4	91.1	94.0	96.2	92.8
NPR	86.7	95.6	77.4	90.8	78.2	97.3	87.7
RINE	92.3	97.1	84.0	96.0	94.9	96.3	93.4
FSD	76.2	97.3	75.3	85.3	71.7	89.3	82.5
Ours	95.1	99.0	91.0	97.3	97.3	95.4	95.9

(b) Generator source attribution.

Method	Excluding Subset						Avg.
	M	G	A	S	V	B	Acc.(%)
UFD	73.6	73.2	69.8	71.7	68.1	73.8	71.7
NPR	84.5	78.4	75.3	82.4	77.1	88.3	81.0
RINE	80.3	77.3	69.5	77.9	71.4	80.0	76.1
FSD	79.5	78.7	71.2	74.6	69.3	79.2	75.4
Ours	91.7	90.8	85.9	87.8	85.0	91.5	88.8

(c) Unseen generator source attribution.

Method	Excluding (test) Subsets						Avg.
	M,G	G,A	A,S	S,V	V,B	B,M	Acc.(%)
UFD	87.1	86.4	69.7	77.4	85.6	82.9	81.5
NPR	85.6	73.9	71.0	62.9	84.8	88.1	77.7
RINE	89.2	84.3	75.8	73.5	82.6	74.7	80.0
FSD	80.0	77.0	60.0	56.2	77.0	64.6	69.1
Ours	91.0	84.6	78.7	82.4	96.2	92.0	87.5

4 EXPERIMENTS

4.1 EXPERIMENT SETTINGS

Datasets and Implementation Details. We evaluate on GenImage (Zhu et al., 2023) dataset, which contains real images and AI-generated images from multiple generative models (Midjourney, Stable Diffusion, ADM, GLIDE, VQDM, and BigGAN). Following the few-shot protocol of FSD (Wu et al., 2025), each generator is treated as a distinct class and all real images form a single real class. To further assess cross-dataset generalization, we additionally evaluate models trained on GenImage using Chameleon (Yan et al., 2025) and Synthbuster (Koutlis & Papadopoulos, 2024) datasets. We use a frozen CLIP ViT-B/16 (Radford et al., 2021; Dosovitskiy et al., 2021) backbone and adopt an episodic training scheme where one generator is held out for testing and the remaining generators are used for training. For fair comparison, UFD (Ojha et al., 2023), NPR (Tan et al., 2024), and RINE (Koutlis & Papadopoulos, 2024) are reproduced under the same prototypical metric learning, based on their official code implementations. All experiments use a 5-shot setting. Additional details including the reproduction of the baseline methods are provided in the appendix.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Binary Classification. We compare our method with various methods under the leave-one-generator-out protocol (Wu et al., 2025). As shown in Tab. 1a, CLIP-based (UFD and RINE) and artifact-oriented detectors (NPR) provide strong baselines. Nevertheless, single-layer CLIP features or simple layer aggregation are insufficient for robust classification. In contrast, our method achieves substantial improvements over both UFD and RINE. This suggests that the advantage of CLRM does not come solely from using CLIP or intermediate features, but from explicitly modeling cross-layer class-patch relations between semantic summaries and spatial evidence.

Generator Source Attribution. Beyond binary classification, we evaluate a more practical setting: generator source attribution, where the model identifies the exact generative source of each image. This experiment encompasses both seen and unseen generators, thereby jointly assessing the model's generalization and source-level discrimination capabilities. This evaluation is particularly important for real-world deployment, where identifying the provenance of an AI-generated image can be as critical as determining whether it is AI-generated. As shown in Tab. 1b, UFD exhibits 71.7% average accuracy, indicating that final layer CLIP embeddings alone are insufficient for source-level discrimination. RINE improves upon UFD by leveraging intermediate CLIP representations, achieving 76.1%; however, this result suggests that global multi-layer descriptors still miss important spatial relational cues. In contrast, our CLRM achieves 88.8% average accuracy, outperforming the strongest baseline NPR by 7.8%p and RINE by 12.7%p. By explicitly modeling cross-layer class-patch relations, our framework captures subtle generator-specific patterns while maintaining strong generalization. This balance between discriminability and transferability enables superior performance.

Table 2: Evaluation under unseen perturbations on the GenImage dataset, applying Gaussian noise (σ=0.1), Gaussian blur (σ=1.0), and WebP compression (QF=80) at test time. Avg. denotes the average over the three perturbation settings.

	Binary Classification					Generator Source Attribution				
	Original	Noise	Blur	Comp.	Avg.	Original	Noise	Blur	Comp.	Avg.
UFD	92.8	69.5	76.3	71.7	72.5	71.7	40.4	50.2	43.2	44.6
NPR	87.7	65.2	74.0	65.5	68.2	81.0	21.5	46.6	38.2	35.4
RINE	93.4	64.5	74.0	67.9	68.8	76.1	36.2	49.6	40.3	42.0
FSD	82.5	70.1	81.1	63.3	71.5	75.4	36.3	57.0	43.3	45.5
Ours	95.9	72.1	85.4	70.9	76.1	88.8	45.0	70.4	58.1	57.8

Unseen Generator Source Attribution. To further evaluate generalization, we extend the leave-one-generator-out setting by excluding two generator subsets during training. The model is trained on the remaining subsets and evaluated on the two unseen generators and the real class (i.e., three-way classification). This setup allows us to examine how effectively the model adapts when multiple unseen generation sources are encountered at test time. As shown in Tab. 1c, our CLRM achieves the best average accuracy, outperforming other baselines. Although some baselines are competitive on specific generator pairs, their performance is less stable across configurations. The explicit relational modeling of cross-layer interactions enables clearer separation among multiple unseen generators while maintaining reliable discrimination from real images.

Open-Set Test Scenario. We assess cross-dataset generalization on Chameleon (CHA) (Yan et al., 2025) and Synthbuster-S (SYN) (Koutlis & Papadopoulos, 2024) datasets (detailed in the appendix) using the models trained on GenImage subsets, without additional fine-tuning. As shown in Tab. 3, ours outperforms competing approaches especially on Chameleon, a challenging dataset that has passed a human perception 'Turing Test'. These results indicate that modeling cross-layer relations yields representations that are robust to generator and dataset shifts, enabling effective generalization without sacrificing source-level discriminability. This supports the potential applicability of CLRM to scenarios in which unseen generators continually emerge.

Table 3: Accuracy on open-set datasets.

Method	CHA	SYN
UFD	65.0	88.3
NPR	55.5	88.1
RINE	80.0	91.3
FSD	58.4	82.5
Ours	82.0	93.0

5 Further Analysis

5.1 Robustness Analysis

Noisy support set. In real-world few-shot deployment, support labels may not always be reliable. For example, a real image may be mistakenly annotated as AI-generated, or a generated image may be mislabeled as real when the source is uncertain. To simulate this, we evaluate all methods under the 5-shot binary classification setting by corrupting the support set at test time. Specifically, we randomly select one real support sample and one fake support sample and swap their labels. This produces one mislabeled example in each class, corresponding to 20% label noise per support class. As shown in Tab. 4, all methods degrade under support-label noise, but ours still achieves the highest accuracy, outperforming the baselines. These results indicate that cross-layer relational embeddings remain highly discriminative even when class prototypes are computed from partially mislabeled support samples.

Table 4: Evaluation on GenImage with noisy support set.

	M	G	A	S	V	B	Avg.
UFD	81.9	98.3	86.6	87.8	91.7	94.4	90.1
NPR	80.2	93.2	73.2	88.7	73.0	97.3	84.3
RINE	91.7	96.3	80.2	95.7	95.1	96.1	92.5
FSD	72.1	96.7	69.7	81.5	67.9	84.5	78.7
Ours	90.1	98.2	88.3	94.8	97.0	95.3	94.0

Unseen Image Perturbation. To evaluate robustness against unseen image perturbations, we apply three distortions to the test images: Gaussian noise, Gaussian blur, and WebP compression. All experiments are conducted on the GenImage dataset under the same protocol as in Tab. 1. We use strong perturbation settings: Gaussian noise with σ=0.1, Gaussian blur with σ=1.0, and WebP compression with a quality factor of 80. The results are summarized in Tab. 2. As expected, all methods experience performance degradation under strong input distortions. The degradation is particularly severe for generator source attribution because low-level perturbations can corrupt fine-grained generator-specific traces. Nevertheless, in binary classification, CLRM achieves the best

performance under Gaussian noise and Gaussian blur, and remains competitive under WebP compression. More importantly, in generator source attribution, CLRM consistently ranks first across all perturbation types. These results suggest that CLRM preserves discriminative inter-generator characteristics by modeling cross-layer semantic-spatial correspondence, rather than relying solely on fragile low-level artifacts that can be easily disrupted by image perturbations.

5.2 Ablation Study on Layer Selection

We investigate how the choice of intermediate layer affects CLRM by varying the layer index and evaluating both binary classification and generator source attribution on GenImage (Fig. 5). The results show a middle-layer preference, with the 5th and 6th layers achieving high performance for both tasks. This behavior reflects a trade-off between semantic alignment and cross-layer complementarity. Earlier layers retain low-level cues that are less aligned with the final semantic representation, while later layers become increasingly similar to the final representation, reducing the complementary information available for cross-layer modeling.

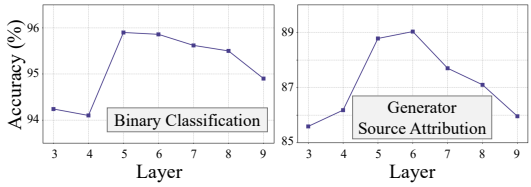


Figure 5: Ablation study on layer selection.

5.3 Generalization to Other Vision Backbones

To investigate whether the effectiveness of CLRM depends on the vision-language pretraining of CLIP, we evaluate our method using two self-supervised vision backbones: DINO (Caron et al., 2021) and DINOv2 (Oquab et al., 2024). Specifically, we use DINO ViT-B/16 and DINOv2 ViT-B/14 under the same leave-one-generator-out 5-shot protocol. In all experiments, the pretrained backbone is kept frozen. As shown in Tab. 5, applying CLRM substantially improves the average performance of both self-supervised backbones across both tasks. These improvements demonstrate that the benefit of CLRM is not restricted to CLIP representations. The substantial gains in generator source attribution also indicate that the final layer class token alone provides limited information for fine-grained source discrimination, whereas cross-layer class-patch relations capture complementary generator-specific characteristics distributed across the backbone hierarchy. These results suggest that cross-layer relational patterns remain effective even when different pretrained vision backbones are employed.

Table 5: Experimental results with DINO and DINOv2 backbones. +Linear applies a linear projection to the final-layer class token, and the resulting embedding is optimized using the same Prototypical Network objective.

	M	G	A	S	V	B	Avg.
Binary Classification							
DINO+Linear	71.5	87.7	74.7	72.3	70.6	93.8	78.4
DINO+Ours	75.6	95.4	82.7	77.4	76.6	94.6	83.7
DINOv2+Linear	71.4	81.8	66.1	70.1	81.3	89.2	76.7
DINOv2+Ours	70.9	96.6	84.7	80.2	92.5	97.0	87.0
Generator Source Attribution							
DINO+Linear	65.7	60.5	62.3	60.7	60.2	67.0	62.7
DINO+Ours	80.1	76.8	77.0	75.9	73.6	82.5	77.6
DINOv2+Linear	55.7	50.8	51.5	53.0	50.4	50.6	52.0
DINOv2+Ours	84.4	81.8	83.6	81.2	83.7	88.9	83.9

6 Conclusion

We addressed few-shot AI-generated image detection under generator shifts, where only limited labeled examples are available for unseen generators. Our analysis shows that real images and images from different generators induce distinct cross-layer behaviors in CLIP, including token transitions and class-patch relational patterns. Based on this observation, we proposed the Cross-Layer Relational Module (CLRM) that models how semantic class tokens correspond to spatial patch evidence between intermediate and final layers. Integrated with prototypical metric learning, CLRM produces transferable metric-space embeddings for few-shot binary classification and generator source attribution. Various experiments show that CLRM outperforms strong baselines on average. These indicate that cross-layer semantic-spatial correspondence provides a useful and generalizable cue for detecting AI-generated images under limited supervision.

AI USE STATEMENT

We used an LLM-based writing assistant solely for language editing, including grammar correction, phrasing refinement, and improving clarity. The tool was not used to generate research ideas, analyses, experiments, or results. All technical content was developed and verified by the authors, who take full responsibility for the manuscript. No proprietary or confidential data beyond the manuscript text was provided to the tool.

ETHICS STATEMENT

This work aims to improve the detection and source attribution of AI-generated images, which can support content authenticity and provenance verification. However, detection errors may lead to incorrect attribution or misclassification; therefore, such systems should be used with appropriate caution, particularly in high-stakes applications.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architecture, training and evaluation protocols, datasets, preprocessing, hyperparameters, and baseline implementations in the main paper and appendix. We also report results across multiple random seeds and provide the computational setup used for our experiments.

REFERENCES

Associated Press. Deepfake of principal's voice is the latest case of ai being used for harm, May 2024. URL https://apnews.com/article/663d5bc0714a3af221392cc6f1af985e.

Lital Binyamin, Yoad Tewel, Hilit Segev, Eran Hirsch, Royi Rassin, and Gal Chechik. Make it count: Text-to-image generation with an accurate number of objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13242–13251, 2025.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

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.

CBS News. Complaints about deepfake ai videos more than doubled this year, fbi says ..., September 2025. URL https://www.cbsnews.com/newyork/news/deepfake-artificial-intelligence-videos/.

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

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

Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multimodal language models see better when they look shallower. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6688–6706, 2025.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 973–982, 2023.

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

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5Ca9sSzuDp.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10696–10706, 2022.

Fusheng Hao, Fengxiang He, Liu Liu, Fuxiang Wu, Dacheng Tao, and Jun Cheng. Class-aware patch embedding adaptation for few-shot image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 18905–18915, 2023.

Jochen Hartmann, Yannick Exner, and Samuel Domdey. The power of generative marketing: Can generative ai create superhuman visual marketing content? *International Journal of Research in Marketing*, 42(1):13–31, 2025. ISSN 0167-8116. doi: https://doi.org/10.1016/j.ijresmar.2024.09.002. URL https://www.sciencedirect.com/science/article/pii/S0167811624000843.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Zexi Jia, Chuanwei Huang, Yeshuang Zhu, Hongyan Fei, Xiaoyue Duan, Zhiqiang Yuan, Ying Deng, Jiapei Zhang, Jinchao Zhang, and Jie Zhou. Secret lies in color: Enhancing ai-generated images detection with color distribution analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13445–13454, 2025.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on computer vision*, pp. 394–411. Springer, 2024.

SuBeen Lee, WonJun Moon, and Jae-Pil Heo. Task discrepancy maximization for fine-grained few-shot classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5331–5340, 2022.

Yunheng Li, Zhong-Yu Li, Quan-Sheng Zeng, Qibin Hou, and Ming-Ming Cheng. Cascade-clip: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. In *International Conference on Machine Learning*, 2024.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8060–8069, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17006–17015, 2024.

Midjourney. Midjourney. `https://www.midjourney.com/home/`, 2022. Accessed: November 11, 2025.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning, PMLR*, 2022.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pp. 86–103. Springer, 2020.

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

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9130–9140, 2024.

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.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. In *International conference on machine learning*, pp. 30105–30118. PMLR, 2023.

Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4136–4145, 2020.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12105–12114, 2023.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28130–28139, 2024.

Jiaan Wang, Sirui Liu, Yu Li, Kaiyuan Yang, Juan Cao, and Sheng Tang. Fleet: Few shots lead effective ai-generated image detection. *arXiv preprint arXiv:2606.31082*, 2026.

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.

Shiyu Wu, Jing Liu, Jing Li, and Yequan Wang. Few-shot learner generalizes across ai-generated image detection. In *Forty-second International Conference on Machine Learning*, 2025.

Wukong. Wukong. `https://xihe.mindspore.cn/modelzoo/wukong`, 2022. Accessed: November 11, 2025.

Jiangtao Xie, Fei Long, Jiaming Lv, Qilong Wang, and Peihua Li. Joint distribution matters: Deep brownian distance covariance for few-shot classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7972–7981, 2022.

Yahoo Finance. A fake livestream featuring an ai-generated, crypto-promoting jensen huang was broadcast at the same time as the nvidia gtc keynote, 2025. URL `https://finance.yahoo.com/news/fake-livestream-featuring-ai-generated-141544943.html`.

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

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6. IEEE, 2019.

Xiaolei Zhao and Xin Zhao. Application of generative artificial intelligence in film image production. *Computer-Aided Design and Applications*, pp. 15–28, 05 2024. doi: 10.14733/cadaps.2024.S27.15-28.

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

A Experimental Settings

A.1 Datasets.

We evaluate our method primarily on the widely used GenImage (Zhu et al., 2023) dataset, which contains real and AI-generated images from diverse generative models. The GenImage dataset includes real images sampled from ImageNet (Deng et al., 2009) and AI-generated images produced by 8 generative models: 7 diffusion models and one GAN. The diffusion-based generators include Midjourney (Midjourney, 2022), Stable Diffusion v1.4 (Rombach et al., 2022), Stable Diffusion v1.5 (Rombach et al., 2022), Wukong (Wukong, 2022), ADM (Dhariwal & Nichol, 2021), GLIDE (Nichol et al., 2022), and VQDM (Gu et al., 2022), while the GAN-based generator is BigGAN (Brock et al., 2019). We treat AI-generated images produced by different generators as distinct classes, while grouping all real images into a single real class. We further observed that Stable Diffusion and Wukong share nearly identical architectures and exhibit highly similar characteristics. Hence, we merge their subsets into a single unified class, referred to simply as Stable

Diffusion (SD). After this adjustment, our dataset comprises images from 7 classes in total: 6 corresponding to different generative models and one representing real images.

To further assess the cross-dataset generalization capability of our method, we additionally evaluate the model trained on GenImage on Chameleon (Yan et al., 2025) and Synthbuster (Koutlis & Papadopoulos, 2024) datasets. This setup also ensures a clear separation between training and test distributions in terms of generative sources. To enforce a stricter evaluation protocol, we remove subsets in Synthbuster that overlap with generators included in GenImage (retain only DALLE-2, DALLE-3, Firefly, Stable Diffusion 2, and Stable Diffusion XL). We denote this filtered version as Synthbuster-S.

A.2 Implementation Details.

We adopt the ViT-B/16 (Dosovitskiy et al., 2021) CLIP image encoder (Radford et al., 2021) as our backbone, keeping it frozen throughout training. All images are resized to 256×256, followed by a random crop of 224×224 and random horizontal flip for data augmentation. During evaluation, a center crop of 224×224 is applied. We extract the 5th and final-layer features of CLIP-B/16. For all reproduced methods, including UFD (Ojha et al., 2023), NPR (Tan et al., 2024), and RINE (Koutlis & Papadopoulos, 2024), we apply the input normalization appropriate for each backbone: CLIP-based models use the official CLIP normalization, while ResNet-50-based models use the standard ImageNet normalization. For each leave-one-generator-out split, one generative model is held out as the test domain, while the remaining 5 generator classes are used for training together with the real class. Within each episode, we randomly sample $N = 3$ classes, each containing $K = 5$ samples, to form both the support and query sets. Note that the experimental results for UFD, NPR, RINE, FSD, and our method in Tab. 1 are averaged over five random seeds. We use AdamW optimizer (Loshchilov & Hutter, 2019), learning rate 3×10^{-4}, and weight decay 0.01 for 10,000 iterations with an episodic batch size of 16. During evaluation, we perform two types of experiments: (1) binary classification, which determines whether an image is real or AI-generated (i.e., $\tilde{N}$=2), and (2) generator source attribution, which identifies the specific generative model responsible for each image. We adopt 5-shot configuration ($\tilde{K}$=5) at test time. Training CLRM for one split takes approximately 2.5 hours on a single NVIDIA RTX A6000 GPU.

A.3 Implementation Details for the Baselines

For UFD (Ojha et al., 2023), NPR (Tan et al., 2024), and RINE (Koutlis & Papadopoulos, 2024), we build upon their official implementations and retain their original input processing and representation extraction procedures. We adapt them to our evaluation setting by replacing their original classification stage with the same Prototypical Network (Snell et al., 2017) framework used for our method. No other method-specific architectural modifications are introduced.

UFD. UFD (Ojha et al., 2023) employs the representation space of a pretrained CLIP image encoder without learning a task-specific visual backbone. Its original formulation performs real-fake classification using a simple classifier on top of the global CLIP image representation. In our implementation, we retain the official CLIP-based feature extraction pipeline and use the final representation as the embedding for prototypical learning. The original classification head is replaced with prototype-based classification, while the CLIP image encoder remains frozen.

NPR. NPR (Tan et al., 2024) is motivated by the observation that upsampling operations in generative models induce characteristic local dependencies among neighboring pixels. It constructs a Neighboring Pixel Relationship representation that emphasizes these generalized local structural artifacts and processes it using its original CNN-based detector. We retain both the NPR input representation and its official feature extractor. The feature produced before the original binary classification layer is used as the embedding for prototypical learning, and only the final classification stage is replaced with the prototype-based classifier.

RINE. RINE (Koutlis & Papadopoulos, 2024) exploits the class-token representations extracted from multiple intermediate Transformer blocks of the CLIP image encoder. The block-wise representations are mapped into a learnable forgery-aware feature space and combined using a Trainable

Table 6: Few-shot binary classification performance in terms of average precision (AP) on the GenImage dataset.

Method	Excluding Subset						Avg.
	Midjourney	GLIDE	ADM	SD	VQDM	BigGAN	
UFD	90.1	99.2	94.1	94.4	96.7	98.1	95.5
NPR	90.2	97.8	81.1	94.5	83.7	98.4	91.0
RINE	96.4	99.0	88.8	98.3	98.0	98.3	96.5
FSD	79.4	98.6	79.0	90.4	74.9	93.2	85.9
Ours	97.3	99.6	94.5	98.6	98.8	98.3	97.9

Table 7: Performance comparison of different architectural variants.

Method	Excluding Subset						Avg.
	Midjourney	GLIDE	ADM	SD	VQDM	BigGAN	Acc.(%)
Binary Classification							
(a) Final Layer	86.7	98.6	90.6	91.1	94.0	96.2	92.8
(b) Direct Difference	88.6	98.3	87.0	94.1	96.1	96.7	93.5
(c) Cross-Layer Implicit Modeling	87.8	97.9	87.0	94.3	96.2	94.1	92.9
(d) Final-Layer Attention Modeling	90.3	98.4	89.6	94.6	96.6	97.7	94.5
(e) Uni-directional Relation Modeling	90.5	98.9	91.3	95.9	96.3	91.8	94.1
(f) Ours	95.1	99.0	91.0	97.3	97.3	95.4	95.9
Generator Source Attribution							
(a) Final Layer	73.6	73.2	69.8	71.7	68.1	73.8	71.7
(b) Direct Difference	81.2	79.8	75.1	77.7	74.6	80.7	78.2
(c) Cross-Layer Implicit Modeling	85.7	83.0	77.0	83.4	79.7	84.2	82.2
(d) Final-Layer Attention Modeling	84.7	83.5	77.3	80.3	77.9	87.5	81.9
(e) Uni-directional Relation Modeling	88.5	87.1	81.4	85.3	81.8	87.8	85.3
(f) Ours	91.7	90.8	85.9	87.8	85.0	91.5	88.8

Importance Estimator, which assigns an importance score to each Transformer block. We retain this intermediate-feature extraction, projection, and importance-weighted aggregation procedure and use the resulting representation as the embedding for prototypical learning.

The official RINE implementation searches over multiple configurations of its projection network, including different numbers of projection layers and projection dimensions. Following this procedure, we also evaluated the corresponding configurations under our few-shot protocol and selected the best-performing one. The configuration used for the results in the main tables consists of a single projection layer with a projection dimension of 512. Apart from replacing the original classification stage with prototypical learning, all remaining components follow the official implementation.

B Additional Experimental Results

B.1 Performance with Average Precision

Tab. 6 provides additional results corresponding to Tab. 1 in the main manuscript. We report average precision (AP) for each test scenario to offer a more comprehensive assessment of detection performance. Consistent with the findings in the main manuscript, our method achieves superior performance across both metrics in most evaluation settings.

B.2 Architectural Variants

In the main manuscript (Sec. 3.5), we analyze the design rationale of CLRM through architectural ablations, based on the few-shot classification performance. Here, we provide detailed results for each excluded generator subset to further validate our design.

Tab. 7 presents the full comparison of architectural variants (a)–(f) across both binary classification and generator source attribution tasks. The variants progressively incorporate the key motivations described in Sec. 3.5: (a) reproduces UFD (Ojha et al., 2023), (b) directly encodes class-token dif-

Table 8: Experimental results with CLIP-L/14. M, G, A, S, V, and B denote Midjourney, GLIDE, ADM, SD, VQDM, and BigGAN, respectively.

	M	G	A	S	V	B	Avg.
Binary Classification							
UFD	88.3	99.6	90.9	95.2	96.4	98.6	94.8
RINE	92.2	98.1	90.0	98.5	96.5	96.4	95.3
Ours	91.9	99.3	92.9	98.2	98.2	98.1	96.4
Generator Source Attribution							
UFD	80.9	79.9	76.5	79.0	75.6	81.6	78.9
RINE	85.7	81.4	77.6	84.5	76.7	85.6	81.9
Ours	92.4	93.0	88.4	91.0	88.7	94.4	91.3

ferences, (c) implicitly models cross-layer relations via concatenation and MLP, (d) models attention within the final layer, (e) explicitly models a cross-layer relation, and (f) employs our full bidirectional relational modeling. The variants correspond to the progression described in Fig. 3. Consistent with the trends observed in Fig. 4, performance generally improves the variants more explicitly model cross-layer relations. The consistent improvements across both binary and attribution settings further validate the effectiveness of CLRM in capturing generator-induced discrepancies.

B.3 Additional Results with CLIP ViT-L/14

To verify whether the effectiveness of CLRM holds with a stronger CLIP backbone, we evaluate CLIP-based methods using CLIP ViT-L/14 under the same leave-one-generator-out 5-shot protocol on GenImage. We compare UFD, RINE, and our CLRM. For ViT-L/14, we use the 12th layer as the intermediate layer, following the same intermediate-layer principle used for ViT-B/16, without additional tuning. As shown in Tab. 8, the stronger backbone improves the overall performance of CLIP-based detectors, but CLRM still achieves the best average performance. The advantage is especially clear in generator source attribution, indicating that even stronger final-layer CLIP embeddings or intermediate class token aggregation remain limited for fine-grained source-level discrimination. Overall, these results show that CLRM scales effectively to a stronger vision-language backbone.

C Limitations

Although our CLRM improves robustness over competing methods, generator source attribution remains sensitive to strong low-level distortions such as noise, blur, and compression. These transformations can corrupt fine-grained generator-specific traces, making attribution substantially harder than binary real-fake classification. This limitation is shared by existing detectors and suggests that robust source attribution under realistic post-processing remains an open challenge.

In addition, like other prototype-based few-shot approaches, our method assumes that a small labeled support set is available for the classes considered at test time. In practice, support annotations may be uncertain or incorrect, and a query image may come from a generator that is not represented in the support set. Although our noisy support experiment shows that CLRM remains robust under a controlled real-fake label-swap scenario, this does not fully address more challenging cases that can occur in real-world scenarios. Addressing these settings is an important direction for future work.