

Rethinking Cross-Generator Image Forgery Detection through DINOv3

Zhenglin Huang¹ Jason Li² Haiquan Wen¹ Tianxiao Li¹ Xi Yang³
Lu Qi⁴ Bei Peng⁵ Xiaowei Huang¹ Ming-Hsuan Yang⁴ Guangliang Cheng¹ [†]

¹University of Liverpool, UK ²Nanyang Technological University

³ HKUST ⁴ UC Merced ⁵ University of Sheffield

Github Page: <https://github.com/hzlsaber/FGTS/>

[†] Corresponding author. E-mail: guangliang.cheng@liverpool.ac.uk

Abstract

As generative models become increasingly diverse and powerful, cross-generator detection has emerged as a new challenge. Existing detection methods often memorize artifacts of specific synthesis models rather than learning transferable cues, leading to substantial failures on unseen generators. Thus, the detection models must have strong generalization to meet task-specific adaptation. Surprisingly, in this work, we find that frozen visual foundation models—particularly DINOv3—already act as good cross-generator detectors, achieving strong performance even *without* any fine-tuning. Through systematic studies across frequency, spatial, and token perspectives, we identify representational trends, suggesting that DINOv3 relies on global, low-frequency structures. These features serve as weak yet transferable **authenticity cues**, rather than on high-frequency, generator-specific artifacts. Motivated by these, we introduce a simple yet generalizable training-free token-ranking strategy followed by a lightweight linear probe. We select only a small subset of authenticity-relevant tokens that consistently improve detection accuracy across all evaluated datasets. Our study provides empirical evidence and a feasible hypothesis for why foundation models generalize across diverse generators. As a result, our model builds a universal, efficient, and interpretable baseline for image forgery detection.

1. Introduction

The rapid evolution of generative AI has fundamentally reshaped visual content creation. High-fidelity images synthesized by diffusion and adversarial models [9, 19, 20, 45] are now virtually indistinguishable from real photographs, posing significant challenges to visual authenticity verification. Despite extensive progress, existing detectors still exhibit poor cross-generator generalization: they perform

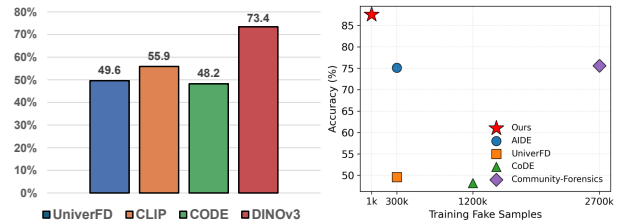


Figure 1. **Cross-generator performance comparison.** **Left:** Average OOD accuracy across the ten commercial generators in So-Fake-OOD [22]. **Right:** OOD detection accuracy versus the amount of training fake data for foundation model-based approaches.

well on generators seen during training but fail catastrophically on unseen ones. This persistent gap raises a central question in modern forgery detection: *How can we develop a detector that generalizes across unseen generators?*

Recent advances in generalized image forgery detection have increasingly focused on model-centric approaches, leveraging powerful pre-trained backbones to improve cross-generator generalization [5, 7, 29, 36, 57]. However, most of these methods remain heavily reliant on large-scale, task-specific adaptation. UniverFD [36] fine-tunes a CLIP backbone [40] on more than 300k samples, and CoDE [5] trains an entirely new contrastive space using millions of diffusion-generated images. As shown in Fig. 1 (left), both methods suffer substantial degradation on unseen commercial generators in the So-Fake-OOD benchmark [22].

This phenomenon raises a fundamental question: *Does the large-scale adaptation process itself introduce bias, or are the foundation models inherently limited?* To investigate this, we compare UniverFD directly against its frozen foundation backbone (CLIP). Remarkably, as shown in Fig. 1 (left), the frozen CLIP model achieves 55.9% accuracy on So-Fake-OOD, outperforming its adapted counterpart by 6.3%. This counter-intuitive finding suggests a critical limitation: large-scale adaptation with hundreds of

thousands of generated images is not only unnecessary, but may also bias the model’s intrinsic generalization capability by overfitting to the artifacts of specific generators.

Although the frozen CLIP model performs better than its adapted counterpart, its overall detection accuracy remains limited. We posit that this limitation stems from its training objective, which aligns visual and linguistic semantics rather than modeling purely visual regularities [40]. Consequently, such language-guided supervision may overlook visual cues that distinguish real from generated content.

Building on this hypothesis, we next explore whether a foundation model trained purely on visual objectives can inherently generalize better for this task. To this end, we turn to **DINOv3** [46], the latest strong baseline in vision foundation models trained exclusively on large-scale image data without any language supervision. Unlike CLIP, DINOv3 learns from image-only self-distillation, encouraging invariance to transformations while preserving fine-grained visual regularities. Under the same zero shot protocol, the frozen DINOv3 model achieves an average accuracy of **73.4%** on the So-Fake-OOD benchmark (Fig. 1 (left)), substantially surpassing all the task-specific detectors trained with hundreds of thousands of samples. This performance is remarkable since DINOv3 was never trained for authenticity discrimination or any forgery-related objective.

These results naturally raise another central question: *Why does a frozen, vision-only foundation model generalize so well to unseen generators?* To explore this, we analyze DINOv3’s internal representations (Sec. 3) across three dimensions—frequency, spatial structure, and token organization. Our findings converge to a consistent pattern: DINOv3 encodes a low-frequency, globally coherent yet spatially distributed signal that systematically differentiates real images from generated ones. We refer to this emergent pattern as **authenticity cues**, a form of visual regularity that is not tied to any specific generator and that is unevenly distributed across patch tokens rather than concentrated in summary tokens such as CLS or register tokens.

Building on these findings, we introduce **Fisher-Guided Token Selection (FGTS)**, a training-free token ranking framework (Sec. 4) that identifies the patch tokens most strongly associated with authenticity cues. To implement this ranking, FGTS adopts the Fisher score [13, 16], which measures each token’s real/fake separability by comparing between-class differences with within-class variability. Using this measure, we find that a compact subset of patch tokens preserves intrinsic authenticity signals while reducing generator-specific noise.

While FGTS uncovers meaningful separability in the frozen feature space, we further explore whether a small amount of supervision can enhance these compact representations without resorting to large-scale adaptation. To this end, we employ a lightweight linear probe trained on

only **1k** fake images from a single generator—two orders of magnitude fewer than prior methods—which preserves the frozen model’s generalization while achieving state-of-the-art cross-generator performance.

Our main contributions are summarized as follows:

- **Empirical Finding.** We uncover that the frozen visual foundation model **DINOv3** exhibits strong cross-generator generalization, surpassing all large-scale training-based detectors under a training-free protocol.
- **Interpretive Analysis.** Through systematic observations, we identify consistent representational trends suggesting that DINOv3 may encode **authenticity cues** through globally coherent, low-frequency structures rather than local artifacts.
- **Methodological Insight.** Building on these observations, we propose **Fisher-Guided Token Selection (FGTS)**, a simple yet effective framework that isolates authenticity-relevant patch tokens from frozen representations. Using only 1k training fake images, FGTS attains **87.5%** accuracy on So-Fake-OOD (Fig. 1 (right)) and **92.6%** on GenImage [63], demonstrating that compact token-level representations can capture strong cross-generator signals with minimal supervision.

2. Preliminaries

2.1. Problem Formulation

Cross-generator image forgery detection aims to determine whether an image x is real or AI-generated. We define a binary classifier $f_\theta : \mathcal{I} \rightarrow \{\text{real}, \text{fake}\}$, where $\mathcal{I}$ denotes the image space. Images are drawn from either the real distribution p_{real} or a generator-specific fake distribution $p_{\text{fake}}^{(g)}$, where $g \in \mathcal{G}$ is a generative model:

$$x \sim \begin{cases} p_{\text{real}}(x), & x \in \mathcal{I}_{\text{real}}, \\ p_{\text{fake}}^{(g)}(x), & x \in \mathcal{I}_{\text{fake}}^{(g)}, g \in \mathcal{G}. \end{cases} \quad (1)$$

In the cross-generator setting, the detector is trained on images from $\mathcal{G}_{\text{seen}}$ and evaluated on disjoint unseen generators $\mathcal{G}_{\text{unseen}}$, where $\mathcal{G}_{\text{seen}} \cap \mathcal{G}_{\text{unseen}} = \emptyset$. The objective is to learn f_θ that generalizes across the distributional shift between generators in $\mathcal{G}_{\text{seen}}$ and $\mathcal{G}_{\text{unseen}}$.

2.2. Representational Concepts

To characterize the types of information that may influence cross-generator generalization, we analyze the visual cues captured by Vision Transformers [10] along two independent but complementary dimensions.

Frequency Dimension. Visual information can be analyzed in the frequency domain:

- **Low-frequency (LF) components:** encode global structure, smooth color or luminance transitions, over-

all layout, and lighting coherence. These signals often remain stable across generation paradigms [12, 14].

- **High-frequency (HF) components:** encode local details, sharp edges, and fine textures, where generator-specific artifacts frequently manifest, such as checkerboard patterns in GANs [1, 26] or synthesis noise characteristics [6].

Spatial Dimension. Visual signals can also be characterized by spatial scope:

- **Local patterns:** spatially confined content within patches or small neighborhoods, including localized inconsistencies [43].
- **Global structure:** image-level coherence requiring integration across distant regions, including scene layout, long-range dependencies, and perspective or illumination consistency [11, 24, 35].

Together, these two dimensions describe the primary forms of visual information that a model may exploit to distinguish real and generated images. In Sec. 3.1, we empirically examine how frozen foundation models, particularly DINOv3, respond to perturbations along these dimensions, providing insight into the representational basis of cross-generator generalization.

2.3. Foundation Models

Frozen DINOv3 exhibits strong cross-generator generalization compared to adapted vision-language models (Figure 1). We summarize the relevant paradigms and highlight the representational factors that will be examined in Sec. 3.

CLIP: Vision-Language Alignment. CLIP (Contrastive Language-Image Pre-training) [40] aligns visual and textual representations through contrastive learning between paired image and text data. This objective encourages semantic alignment across modalities and emphasizes category-level correspondence rather than fine-grained visual regularities.

DINO: Vision-Only Self-Distillation. The DINO family [4, 37, 46] learns visual representations purely from images through self-distillation, where a student network matches a teacher’s predictions across multiple augmented views to promote invariance while preserving discriminative structure. In this work, we focus on **DINOv3**, a vision-only self-distilled transformer whose architecture includes three token types: CLS, register, and spatial patch tokens. These play distinct functional roles, as summarized below.

CLS Token. A global summarizer trained purely on visual consistency without linguistic supervision as in CLIP.

Register Tokens. Learnable non-spatial tokens designed to store image-level statistics and prevent artifact accumulation in patch tokens [8, 23]. In DINOv2 [37], these tokens can be attached post hoc after training, whereas DINOv3 [46] integrates them natively during training, allowing all token types to be learned jointly from the beginning.

Patch Tokens. Spatially grounded tokens that represent

localized image content and serve as the primary carriers of visual detail.

This architecture provides multiple token types that capture visual information at different levels of abstraction. In Sec. 3.2, we empirically analyze how these representations behave across spatial, frequency, and token-level dimensions to better understand their contribution to cross-generator robustness.

3. Empirical Observations and Analysis

Building on the frequency–spatial framework introduced in Sec. 2, we conduct a series of empirical analyses to uncover how DINOv3 encodes authenticity-related information. Specifically, Sec. 3.1 studies DINOv3 from two complementary perspectives—the **frequency domain** and **spatial perturbations**. Sec. 3.2 then investigates the **token-level mechanisms** that underpin global image representations. Together, these observations motivate the hypothesis and method developed in Sec. 4.

To ensure a fair and representative evaluation of cross-generator generalization, our analyses focus on the **commercial diffusion generators** in the So-Fake-OOD benchmark [22], which encompass ten major proprietary text-to-image systems with diverse architectures and training pipelines. This subset best reflects real-world generative diversity and unseen conditions. Comprehensive quantitative results on other generators, including GANs and open-source diffusion models, are reported in Sec. 5.

3.1. Frequency-Spatial Evidence

Prior work has shown that different generative families tend to introduce artifacts with distinct spectral and structural characteristics. For example, GANs often produce high-frequency patterns such as checkerboard effects [12, 14], while diffusion models exhibit more subtle low-frequency deviations in color or geometry [6]. These findings suggest that frequency and spatial organization provide a natural lens for studying how authenticity-related signals may appear across different generators. Motivated by this perspective, we examine which components along these two dimensions DINOv3 is actually sensitive to when generalizing to unseen models. This consideration leads to two guiding questions. First, does DINOv3 rely more on low-frequency structure or on high-frequency details? Second, are the relevant cues primarily local or globally coherent? The following two empirical observations address these questions.

Observation A: Low-frequency dominance. We first probe DINOv3’s spectral sensitivity by applying ideal frequency domain filters with cutoff ratios $r \in \{0.1, 0.3, 0.5\}$. Under the **low-pass (LP)** condition, only coarse, low-frequency components are preserved, while fine details are removed. The **high-pass (HP)** condition keeps high-frequency signals while discarding global structure. Av-

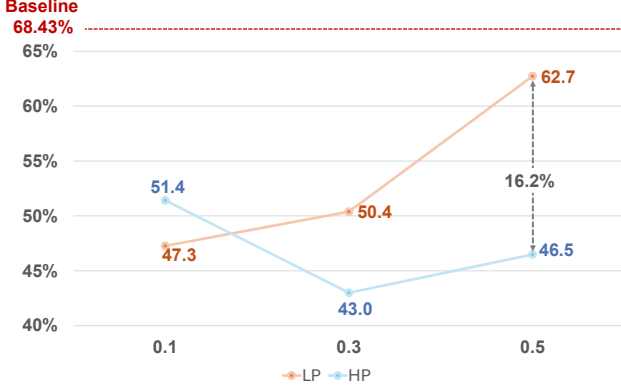


Figure 2. **Low-pass vs. high-pass filtering on DINOv3.** Average accuracy under low-pass (LP) and high-pass (HP) filtering across cutoff ratios on ten diffusion generators.

erage accuracy across ten commercial diffusion generators (Fig. 2) reveals a pronounced asymmetry: As the cutoff increases, LP accuracy steadily recovers, whereas HP accuracy remains almost unchanged and far below the baseline. At $r = 0.5$, the LP-HP gap reaches **16.2%**, confirming that DINOv3’s authenticity representation is strongly biased toward low-frequency information. This tendency likely stems from its self-distillation objective, which promotes invariance to high-frequency variations and emphasizes stable, slowly varying spatial statistics.

Observation B: Dependence on global spatial coherence. While frequency analysis reveals a strong preference for low-frequency information, it remains unclear whether these cues originate from locally smooth textures or truly global structural organization. To clarify this, we apply two complementary spatial perturbations: (1) **Random Masking**, which removes local content by replacing 50% of patches with their mean pixel values while preserving the global layout; and (2) **Patch Shuffling**, which destroys spatial coherence by randomly permuting patch positions within local neighborhoods while preserving local textures.

Fig. 3 visualizes the *accuracy drop difference* (Shuffle-Mask) across ten commercial diffusion generators. DINOv3 shows near invariance to masking (average drop **0.1%**), confirming that local pixel removal has minimal effect. However, once spatial coherence is disrupted, performance decreases sharply (average drop **5.6%**, up to **11.1%** on HiDream). This result demonstrates that DINOv3’s decision boundary critically depends on maintaining global spatial organization rather than on localized details.

However, because frequency and spatial structure are coupled in natural images, these observations still leave open whether the low-frequency preference arises from retained spectral energy or coherent spatial organization.

Observation C: Frequency and spatial interaction. Building on Observations A and B, we next examine

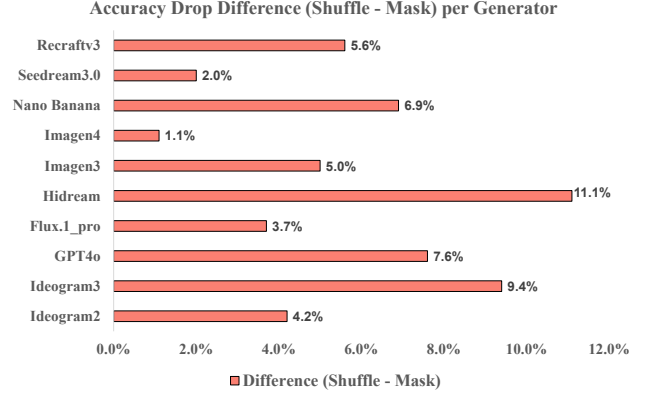


Figure 3. **Impact of spatial coherence on DINOv3.** Accuracy drop difference (ΔAcc , Shuffle-Mask) under 50% perturbation across ten commercial diffusion generators from So-Fake-OOD.

whether the low-frequency preference is driven by preserved spatial structure rather than by retained spectral energy alone. To separate these factors, we conduct a joint frequency and spatial perturbation analysis under three controlled conditions (Fig. 4). **Condition A** applies a global low-pass filter followed by complete patch shuffling, removing all global structural coherence. **Condition B** uses block-wise low pass filtering with intra block shuffling, disrupting coarse spatial alignment while preserving local textures. **Condition C** applies the same block-wise low pass filtering without shuffling as a control.

Across our setting, all metrics exhibit a clear and consistent trend. **Condition A** collapses to near random behavior, indicating that the discriminative signal vanishes once global spatial structure is destroyed. **Condition B** yields only partial recovery, suggesting that isolated local structure provides limited support. **Condition C** stays close to the low pass baseline, confirming that block-wise low pass filtering alone introduces minimal artifacts. Overall, these results demonstrate that DINOv3’s apparent low-frequency dependence is in fact a dependence on spatially coherent global structure rather than on low-frequency energy itself.

Interpretation. Together, the three observations reveal a consistent representational pattern. DINOv3 distinguishes real from generated images through globally coherent low-frequency structures that encode spatial organization and overall scene coherence. Performance remains stable when local details are removed but collapses once global coherence is disrupted, indicating that structural organization rather than local texture drives authenticity perception. This global structural bias explains DINOv3’s strong generalization, since inconsistencies in perspective, illumination, and layout tend to persist across different diffusion models, whereas local artifacts vary widely. However, these findings describe the phenomenon at the image level and leave open a key question: *How is this global authenticity information*

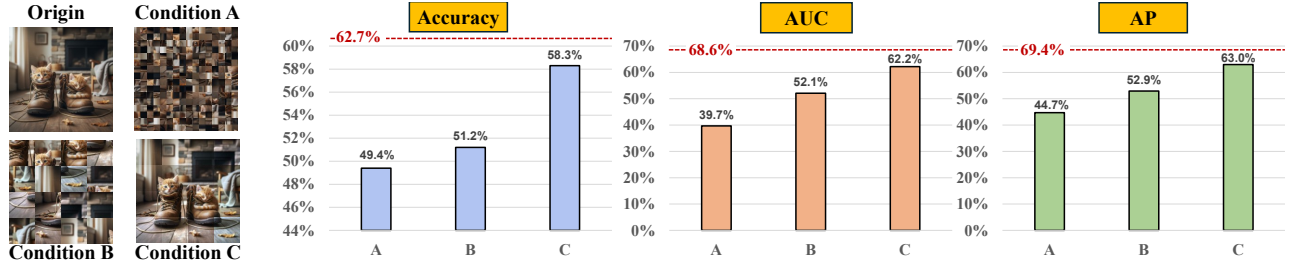


Figure 4. **Performance under spatial disruption conditions.** Left: Visual examples of each condition. Right: Accuracy, AUC, and AP across three experimental conditions. The red dashed line indicates the low-pass only baseline (LP, $r=0.5$) without spatial disruption.

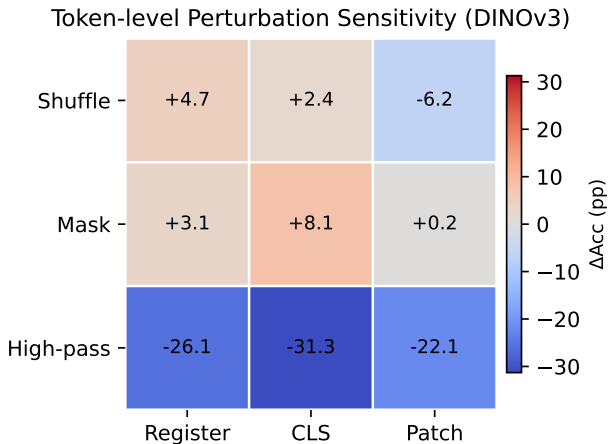


Figure 5. **Token-level perturbation sensitivity in DINOv3.** Heatmap of ΔAcc (percentage points), where $\Delta\text{Acc} = \text{Acc}_{\text{pert}} - \text{Acc}_{\text{base}}$ for each token type under Shuffle, Mask, and High-pass perturbations. Positive values (red) indicate increased accuracy and negative values (blue) indicate decreased accuracy.

encoded in the model, and which tokens contribute to it?

3.2. Where Are Authenticity Cues Encoded?

The previous section showed that DINOv3 distinguishes real from synthetic images by relying on **globally coherent low-frequency structures**. We refer to these structures as **authenticity cues**, a form of generator agnostic signal that reflects the physical and perceptual consistency of an image. Having characterized these cues at the image level, we next ask how they are represented inside DINOv3. Since the architecture combines spatial patch tokens with non-spatial CLS and register tokens (Sec. 2.3), a key question is whether authenticity cues reside primarily in the non-spatial tokens or emerge as a distributed pattern across patches. To address this, we perform controlled perturbation analyses and discriminability evaluations on each token type.

Experimental Protocol. To examine the role of different token types, we extend the perturbation analyses from Sec. 3.1 to the token level. For each token type (CLS, Register, Patch) in DINOv3’s final layer, we (1) extract features

Table 1. **Token-wise evaluation on frozen DINOv3.** Average accuracy (Acc), AUC, and AP across evaluation sets under a linear probe protocol. Patch-only nearly matches using all tokens, while CLS and Register tokens lag behind, indicating that authenticity information is distributed among spatial patches.

Token strategy	Acc	AUC	AP
All (201 tokens)	0.7364	0.8100	0.8228
CLS (1 token)	0.7053	0.8059	0.8199
REG (4 tokens)	0.6827	0.7703	0.7785
Patch (196 tokens)	0.7403	0.8132	0.8262
CLS + REG (5 tokens)	0.7134	0.8002	0.8020
CLS + Patch (197 tokens)	0.7346	0.8070	0.8208

under frequency and spatial perturbations, (2) compute the mean feature of real and synthetic samples for that token type using a small, balanced reference set, and (3) evaluate detection performance by comparing test samples against these mean features. This provides a training-free measure of how well each token type separates real/fake images.

To further assess the discriminative contribution of each token type, we additionally train a lightweight linear classifier on top of features from each token subset using 1,000 real and 1,000 synthetic images from a single generator. This complementary evaluation reveals how linearly separable each token type is under minimal supervision.

Observation D: Distributed representation of global authenticity. Fig. 5 reveals distinct behaviors across token types. Non-spatial tokens (CLS and Register) are largely position independent, showing improved accuracy under shuffling (+2.4%, +4.7%) and substantial degradation under high-pass filtering (−31.3%, −26.1%). This pattern indicates that these tokens rely primarily on low-frequency global statistics rather than on spatial structure. Patch tokens show the opposite trend. Their accuracy decreases when spatial order is disrupted (−6.2%), confirming their dependence on coherent spatial organization, yet they exhibit comparable low-frequency sensitivity under high-pass filtering (−22.1%). Tab. 1 further shows that patch-only

features achieve the highest accuracy (74.0%), surpassing CLS (70.5%) and Register (68.3%) tokens.

Interpretation. These results suggest that DINOv3 distributes global information across both non-spatial and spatial tokens, but with different functional roles. CLS and Register tokens capture broad, position-agnostic low-frequency statistics, whereas patch tokens integrate these low-frequency components with coherent spatial organization. Since authenticity cues hinge on globally coherent low-frequency structure (Sec. 3.2), such spatial grounding becomes particularly important for real/fake discrimination. This explains why patch-only features (74.0%) slightly outperform using all tokens (73.6%): patch tokens carry the most task-relevant portion of DINOv3’s global low-frequency representation.

4. Method

4.1. Motivation

The analyses in Sec. 3.2 show that DINOv3’s cross-generator robustness is driven by globally coherent low-frequency structure distributed across patch tokens. Tab. 1 further indicates that patch tokens collectively provide the strongest discriminative signal, but it remains unclear whether their contributions are uniform or whether certain tokens play a disproportionately important role in encoding authenticity cues.

This motivates a sparsity perspective: global coherence may be concentrated more strongly in a subset of patch tokens rather than evenly distributed. Similar sparsity phenomena have been observed broadly in ViTs, where only a small fraction of tokens carry most of the semantic or task-relevant information [41, 44, 55]. If so, we expect: **(P1)** a compact subset ($K \ll 196$) to match or exceed full-patch performance; **(P2)** selected tokens to exhibit stronger real/fake discrimination; **(P3)** selection patterns to remain stable across unseen generators. These predictions call for a principled way to identify which tokens are informative.

4.2. Rationale for Token Selection

Section 4.1 indicates that authenticity cues, although global in nature, are not uniformly represented across patch tokens. This creates the need for a mechanism that identifies the tokens that most reliably capture low-frequency coherence without modifying the DINOv3 backbone. Such a mechanism should operate directly in the frozen feature space, require minimal supervision, and downweight noisy or weakly informative tokens.

To meet these requirements, This motivates the search for a simple and quantitative criterion that reflects how strongly each token separates real and synthetic samples. We therefore introduce **Fisher-Guided Token Selection (FGTS)**, a training-free procedure that scores patch tokens

by their discriminative stability across real and synthetic distributions. FGTS operationalizes the sparsity perspective by selecting the tokens most reflective of the coherence signals that underpin cross-generator detection.

4.3. The FGTS Framework

FGTS consists of two stages: an **offline token-ranking stage** that scores each patch token by its discriminative reliability, and a **lightweight inference stage** that selects and aggregates a compact subset of tokens for final prediction.

4.3.1. Stage 1: Offline Token Ranking

Given a small, balanced reference set (1000 real and 1000 generated images), we compute for each patch token a scalar score indicating how well its features separate the two classes. Specifically, we adopt the **Fisher Discriminability Ratio**, defined for the i -th token as:

$$F_i = \frac{(\mu_{\text{real},i} - \mu_{\text{fake},i})^2}{\sigma_{\text{real},i}^2 + \sigma_{\text{fake},i}^2}, \quad (2)$$

where $\mu_{\text{real},i}$ and $\sigma_{\text{real},i}^2$ denote the mean and variance of token i for real samples, and likewise for fake ones.

A higher F_i indicates that token i produces features with greater between-class separation and lower within-class variability. We rank all tokens according to F_i and store the sorted index list $\mathcal{I} = [i_1, i_2, \dots, i_N]$ for downstream use. This ranking step requires no gradient updates, is computationally inexpensive, and provides a quantitative estimate of each token’s discriminative stability.

4.3.2. Stage 2: Lightweight Inference

At inference time, FGTS uses only the top- K ranked tokens to form a compact representation. Given an input image:

1. **Extract Features:** Pass the image through a frozen backbone (for example, DINOv3) to obtain N patch tokens $\{p_1, \dots, p_N\}$.
2. **Select Tokens:** Retrieve the subset corresponding to the top- K indices $\mathcal{I}_K = \{i_1, \dots, i_K\}$.
3. **Aggregate:** Average the K selected vectors to obtain an embedding $z_{\text{out}} \in \mathbb{R}^D$.

This yields a compact representation that preserves the most informative portion of the frozen feature space.

4.4. Classification Protocols

(a) Training-Free Protocol. No learnable parameters are introduced. We precompute the centroids of real and fake embeddings, μ_{real} and μ_{fake} , using the same top- K tokens from the reference set. A test image is classified by its cosine similarity to these centroids. This directly measures how separable the frozen representation is.

(b) Linear Probe Protocol. To assess linear separability under minimal supervision, we train a lightweight logistic

Table 2. **Cross-generator generalization on So-Fake-OOD.** All numbers report accuracy (%). SD and PG denote models fine-tuned on SD v1.4 and ProGAN, respectively, and LD denotes the CLIP adapted using LDM-generated images.

Detection method	Flux.1_pro	GPT-4o	HiDream	Ideogram 2	Ideogram 3	Imagen 3	Imagen 4	Recraft-v3	Seedream 3.0	Nano Banana	Avg-acc
Community-Forensics [38]	59.37	86.14	80.48	66.53	77.68	75.36	77.74	<u>75.65</u>	58.59	81.87	<u>75.61</u>
UniverFD [36]	49.58	49.63	65.43	45.29	46.95	45.96	47.15	<u>47.93</u>	49.66	48.18	49.57
CLIP-LD [40]	66.80	69.67	71.13	53.15	66.69	77.42	67.06	73.71	68.38	74.14	68.82
CoDE [5]	48.71	42.36	41.03	43.62	45.18	46.03	45.88	49.66	46.74	44.39	45.37
AIDE-PG [57]	49.91	97.42	56.46	51.31	<u>95.71</u>	94.71	<u>77.89</u>	71.91	59.83	96.12	75.13
AIDE-SD [57]	51.15	88.55	63.21	53.15	90.31	86.22	72.81	68.29	<u>70.71</u>	92.61	73.70
Ours (Training free)	<u>68.90</u>	84.20	<u>83.33</u>	56.96	83.55	79.59	69.90	71.00	67.42	85.77	75.06
Ours (Linear probe)	79.90	<u>96.04</u>	96.02	<u>63.80</u>	95.91	<u>92.83</u>	90.53	86.57	77.87	<u>95.88</u>	87.53

Table 3. **Evaluation on GenImage.** Accuracy (%). The best and second-best results are shown in **bold** and underline, respectively.

Method	Midjourney	SD 1.4	SD 1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	z Avg.
ResNet50 [17]	54.9	<u>99.7</u>	99.7	53.5	61.9	98.2	56.6	52.0	72.1
DeiT-S [50]	55.6	99.9	<u>99.8</u>	49.8	58.1	98.9	56.9	53.5	71.6
Swin-T [31]	62.1	99.9	<u>99.8</u>	49.8	67.6	<u>99.1</u>	62.3	57.6	74.8
CNNSpot [51]	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec [60]	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
F3Net [39]	50.1	99.9	99.9	49.9	50.0	99.9	49.9	49.9	68.7
GramNet [30]	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9
UnivFD [36]	73.2	84.2	84.0	55.2	76.9	75.6	56.9	80.3	73.3
PatchCraft [61]	79.0	89.5	89.3	77.3	78.4	89.3	83.7	72.4	82.3
AIDE [57]	<u>79.4</u>	<u>99.7</u>	<u>99.8</u>	<u>78.5</u>	<u>91.8</u>	98.7	80.3	66.9	86.9
Ours (Training free)	72.3	97.5	89.7	75.9	86.7	95.2	97.6	<u>90.7</u>	<u>88.2</u>
Ours (Linear probe)	85.7	98.1	92.9	90.7	92.0	93.6	<u>93.5</u>	94.4	92.6

regression classifier on the compact representation z_{out} using only **1k real and 1k fake images from a single generator**, while keeping the backbone entirely frozen.

5. Experiments

5.1. Experimental Setting

Baseline Detectors. We compare FGTS with two families of detectors. (1) Traditional detectors, including CNNSpot [51], FreDect [15], LNP [3], Fusing [24], LGrad [48], Spec [60], F3Net [39], DIRE [52], GramNet [30], PatchCraft [61], and NPR [49]. (2) Foundation model based detectors, including AIDE [57], UniverFD [36], CoDE [5], and Community Forensics [38].

Datasets. FGTS is evaluated on three benchmarks, each following its standard reference set: (I) So-Fake-OOD [22] uses 1,000 real and 1,000 fake images from **LDM**; (II) GenImage [63] uses 1,000 real and 1,000 fake images from **Stable Diffusion 1.4**; and (III) AIGCDetectionBenchmark [61] uses 1,000 real and 1,000 fake images from **ProGAN** [25]. These reference sets are shared across both protocols.

Evaluation Protocols. All experiments are conducted using a frozen **DINOv3-ViT-7B** backbone. Two evaluation settings are employed. In the **training-free** setting, we compute real and fake centroids from a balanced reference

set of 1,000 real and 1,000 fake images and classify each test sample by the cosine similarity between its FGTS embedding and these centroids. In the **linear-probe** setting, a lightweight logistic regression classifier is trained on the same reference set while the backbone remains frozen. Unless stated otherwise, all main results use $K = 10$ selected patch tokens, as supported by the ablation in Sec. 5.6.

5.2. Experimental Results on So-Fake-OOD

We evaluate FGTS on the **So-Fake-OOD** benchmark [22], which includes ten commercial diffusion models unseen by all detectors. Tab. 2 reports results for both our training-free and 1k-sample linear probe variants. FGTS already performs competitively in the training-free setting, and the linear probe further improves accuracy to **87.53%**, outperforming all comparison methods on average.

5.3. Experimental Results on GenImage

Tab. 3 reports results on GenImage [63], covering eight representative generators. FGTS achieves the highest average accuracy with the linear probe (**92.6%**), and its training-free variant ranks second, indicating strong cross-generator transfer even without learning.

Table 4. Comparison on the AIGCDetectionBenchmark [61]. Accuracy (%) across detectors and generator categories. All methods are fine-tuned on ProGAN, except for DIRE-D, which follows its official setup and is trained using ADM-generated fake images.

Method	ProGAN	StyleGAN	BigGAN	CycleGAN	StarGAN	GauGAN	StyleGAN2	WFIR	ADM	Glide	Midjourney	SD v1.4	SD v1.5	VQDM	Wukong	DALLE2	Mean
CNNSpot [51]	100.00	90.17	71.17	87.62	94.60	81.42	86.91	91.65	60.39	58.07	51.39	50.57	50.53	56.46	51.03	50.45	70.78
FreDect [15]	99.36	78.02	81.97	78.77	94.62	80.57	66.19	50.75	63.42	54.13	45.87	38.79	39.21	77.80	40.30	34.70	64.03
Fusing [24]	100.00	85.20	77.40	87.00	97.00	77.00	83.30	66.80	49.00	57.20	52.20	51.00	51.40	55.10	51.70	52.80	68.38
LNP [3]	99.67	91.75	77.75	84.10	<u>99.92</u>	75.39	94.64	70.85	84.73	80.52	65.55	85.55	85.67	74.46	82.06	88.75	83.84
LGrad [48]	99.83	91.08	85.62	86.94	99.27	78.46	85.32	55.70	67.15	66.11	65.35	63.02	63.67	72.99	59.55	65.45	75.34
UniverFD [36]	99.81	84.93	95.08	98.33	95.75	99.47	74.96	86.90	66.87	62.46	56.13	63.66	63.49	85.31	70.93	50.75	78.43
DIRE-G [52]	95.19	83.03	70.12	74.19	95.47	67.79	75.31	58.05	75.78	71.75	58.01	49.74	49.83	53.68	54.46	66.48	68.68
DIRE-D [52]	52.75	51.31	49.70	49.58	46.72	51.23	51.72	53.30	98.25	<u>92.42</u>	<u>89.45</u>	91.24	91.63	91.90	90.90	<u>92.45</u>	71.53
PatchCraft [61]	100.00	92.77	<u>95.80</u>	70.17	99.97	71.58	89.55	85.80	82.17	83.79	90.12	95.38	95.30	88.91	91.07	96.60	89.31
NPR [49]	99.79	<u>97.70</u>	84.35	<u>96.10</u>	99.35	82.50	98.38	65.80	69.69	78.36	77.85	78.63	78.89	78.13	76.11	64.90	82.91
AIDE [57]	<u>99.99</u>	99.64	83.95	98.48	99.91	73.25	<u>98.00</u>	<u>94.20</u>	<u>93.43</u>	95.09	77.20	93.00	<u>92.85</u>	<u>95.16</u>	<u>93.55</u>	96.60	92.77
Ours (Training free)	99.25	77.88	75.73	69.84	50.10	<u>99.79</u>	75.71	77.20	59.90	75.10	50.60	74.90	72.90	94.40	80.00	60.50	78.99
Ours (Linear probe)	99.59	89.18	99.18	89.53	94.52	99.80	90.92	98.20	85.76	93.04	77.32	<u>93.52</u>	92.66	98.72	95.60	77.00	<u>92.45</u>

Table 5. Robustness under perturbations on So-Fake-OOD.

Perturbation	Acc	AUC	AP
Clean	0.8750	0.9527	0.9561
Gaussian (5)	0.7354	0.8411	0.8641
Gaussian (10)	0.5531	0.4779	0.5161
JPEG (70)	0.6981	0.7656	0.8154
JPEG (80)	0.7514	0.8612	0.8713
Resize (0.5)	0.8407	0.9225	0.9393
Resize (0.75)	0.8534	0.9349	0.9411

5.4. Results on AIGCDetectionBenchmark.

We also evaluate FGTS on the AIGCDetectionBenchmark [61]. Results are summarized in Tab. 4. Despite relying only on the frozen DINOv3 backbone and a small 1k real / 1k fake reference set, FGTS remains highly competitive. The linear probe achieves an average accuracy of **92.45%**, which is on par with the best-performing methods and clearly exceeds all traditional and most foundation model-based detectors.

5.5. Robustness Experiment

We evaluate the robustness of FGTS under common perturbations including Gaussian noise, JPEG compression, and image downsampling. As summarized in Tab. 5, FGTS remains stable under mild corruptions (JPEG 80, resize 0.75).

5.6. Ablation Experiment

Top-K Token Selection. We assess the effect of selecting different numbers of tokens by comparing FGTS with a Random-K baseline. For each $K \in \{10, 20, 30, 50\}$, FGTS selects the top-K patch tokens ranked by Fisher scores, while Random-K samples tokens uniformly at random. Both use the frozen DINOv3 backbone and the same training free protocol. As shown in Fig. 6, FGTS consistently outperforms random selection across all K, with the largest gain at smaller token budgets (for example, $K = 10$). This indicates that FGTS identifies a more informative subset of patch tokens.

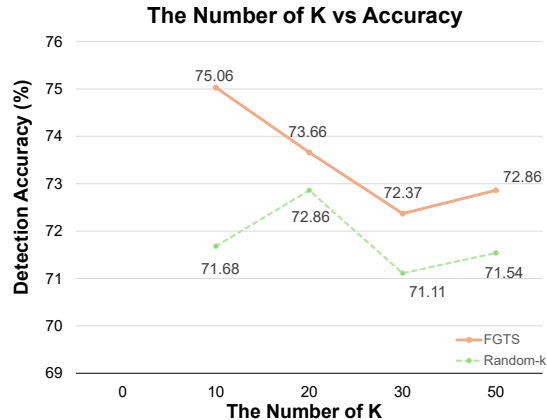


Figure 6. Impact of the number of selected tokens.

Reference Generator. We evaluate how the choice of reference generator affects performance. As shown in Tab. 6, using LDM, SD v1.4, or ProGAN yields comparable accuracy, indicating that FGTS is largely insensitive to reference.

Table 6. Reference generators used for linear probe training on So-Fake-OOD.

Generator	Acc (%)
LDM	87.5
SD v1.4	86.8
ProGAN	84.9

6. Related Work

Image Forgery Detection. Early detectors rely on CNNs that learn generator-specific artifacts [14, 34, 51, 59], while frequency and gradient-based methods [11, 12, 15] improve robustness but still fail on unseen generators. Model-centric approaches such as UniverFD [36] and CoDE [5] rely on large backbones and extensive synthetic training. Recently, large vision-language models have also been explored for forgery detection, showing strong cross-generator robustness [21, 27, 28, 28, 56]. In contrast, FGTS avoids task-specific training entirely and uses the global features of a

frozen DINOv3 to identify a compact subset of tokens for cross-generator detection.

Visual Foundation Models. Large visual foundation models such as CLIP [40, 47, 54], MAE [18, 53], SLIP [33], BEiT [2], and iBOT [62] provide strong transferable features, but their training objectives emphasize semantic alignment or reconstruction and therefore tend to pay less attention to the global low-frequency structures that matter for authenticity analysis. Self-distilled models such as DINOv2 [37] begin to exhibit stronger global invariance, yet their representations remain less cleanly separated across token types. DINOv3 [46], trained at a significantly larger scale and equipped with native register tokens, exhibits noticeably cleaner global feature organization and robust invariances that align naturally with the demands of cross-generator forgery detection.

7. Conclusion

We investigated how DINOv3 encodes real versus synthetic content and found that global low-frequency structure emerges as a transferable **authenticity cue**, distributed across a subset of spatial tokens rather than concentrated in non-spatial representations. Leveraging these observations, we proposed **FGTS**, a training-free token selection strategy that activates a frozen DINOv3 as a universal detector. Our results suggest that representation-centered approaches can provide a lightweight and effective alternative to task-specific training. **Limitations and future work.** This study provides an initial analysis and does not yet offer a complete understanding of global representations in foundation models. Moreover, the analysis centers on DINOv3, and it remains unclear how well the observations transfer to other foundation models or multimodal architectures. In addition, our investigation is limited to still images and does not address temporal consistency or video-specific artifacts. Finally, our evaluation covers only existing generators, which may evolve rapidly. Future work will therefore explore broader model families, temporal modeling, and more general token selection strategies.

8. Appendix

Contents of the Appendices:

Section A: Implementation Details and Computational Efficiency.

Section B: Impact of Model Scaling and Architecture.

Section C: Effectiveness of Fisher-Guided Token Selection (FGTS).

Section D: Analysis of Large-Scale Adaptation.

Section E: Analysis of Performance Variations Across Generators.

A. Implementation Details and Computational Efficiency

This section provides the full experimental configurations used in all evaluations, followed by a detailed comparison of computational efficiency across baseline detectors.

A.1. Experimental Settings

Backbone Configuration. All experiments employ the **DINOv3-ViT-7B** [46] model as a frozen feature extractor. Images are resized to 224×224 and normalized following the official preprocessing pipeline. Features are taken from the final transformer block without any form of test-time augmentation, fine-tuning, or distillation.

Reference Set Construction. To ensure a fair comparison with fully supervised detectors, we match the exact domain assumptions adopted in prior works. Most supervised baselines are trained using a single synthetic source per benchmark; therefore, we construct a compact reference set of **1,000 real** and **1,000 synthetic** images that mirrors their training setup:

- **Real:** Following CNNSpot [51], we sample from one LSUN [58] category (e.g., `car`).
- **Synthetic:** For each benchmark, we select the same generator family that supervised baselines predominantly train on:
 - AIGCDetectionBenchmark: **ProGAN** [25].
 - So-Fake-OOD: **Latent Diffusion (LDM)** [42].
 - GenImage: **Stable Diffusion v1.5**.

This alignment removes domain discrepancies and ensures that our linear probe is evaluated under the same generative-source assumptions as the supervised detectors.

Linear Probe. For settings requiring supervision, we train a minimal linear classifier on top of frozen DINOv3 features:

- **Architecture:** A single fully connected layer ($4096 \rightarrow 2$).
- **Optimization:** Adam ($\text{lr}=1 \times 10^{-2}$), cross-entropy loss.
- **Schedule:** 50 epochs, batch size 32.

Since the backbone is frozen, training is extremely lightweight; the dominant cost of our pipeline lies in a one-time feature extraction pass over the 2k reference images.

A.2. Computational Efficiency

A major advantage of our framework is the elimination of backbone fine-tuning, which constitutes the bulk of training cost in existing supervised detectors. We compare with widely used baselines using different backbone architectures: **CNNSpot** [51] (ResNet-50 [17]), **Uni-verFD** [36] (CLIP-ViT-L [40]), **CoDE** [5] (CLIP-ViT-T), **Community-Forensics** [38] (CLIP-ViT-L), and **AIDE** [57] (CLIP-ConvNeXt [32]).

Cost Components. Our compute overhead consists of:

Table 7. **Computational Cost Comparison.** For our Training free variant, the 2k images are used only to construct the reference set and do not involve parameter updates. “-” in the GPU setup column indicates that the original paper provides no hardware details. Speedup is measured relative to CNNSpot.

Method	Backbone	Training Data	Trainable Params	Training Time	GPU Setup	Speedup
CNNSpot [51]	ResNet-50	~720k	~25M	~24h	–	1×
UniverFD [36]	CLIP-ViT-L	~720k	~1.5k	~12h	–	2×
CoDE [5]	CLIP-ViT-T	~12M	~5M	~48h	RTX6000×4	0.5×
Community-Forensics [38]	CLIP-ViT-L	~5.4M	~22M	~72h	–	~0.3×
AIDE [57]	CLIP-ConvNeXt	~300k	~10M	~2h	A100×8	12×
Ours (Linear probe)	DINOv3-ViT-7B	2k	8,194	<5 min	RTX5090×1	~300×
Ours (Training free)	DINOv3-ViT-7B	2k	0	0	RTX5090×1	N/A

Table 8. **Impact of model scaling on cross-generator generalization.** We evaluate five DINOv3 and four DINOv2 models of increasing capacity on the So-Fake-OOD benchmark. All numbers report accuracy (%). Larger DINOv3 models exhibit clear scaling behavior, while DINOv2 shows weaker improvements under the same evaluation protocol.

Model	Flux.1_pro	GPT-4o	HiDream	Ideogram 2	Ideogram 3	Imagen 3	Imagen 4	Recraft-v3	Seedream 3.0	Nano Banana	Avg-acc
DINOv3											
DINOv3-S/16	55.6	67.1	69.9	54.2	74.4	60.5	68.5	70.3	58.3	67.0	64.6
DINOv3-B/16	62.3	76.0	81.0	51.4	79.7	72.8	70.8	70.3	64.3	74.5	70.3
DINOv3-L/16	71.9	82.5	86.6	58.1	88.9	76.6	77.7	73.2	68.2	81.8	76.7
DINOv3-H/16	76.5	85.0	87.7	61.2	88.3	78.5	73.7	72.6	69.9	84.7	77.8
DINOv3-7B	79.9	96.0	96.0	63.8	95.9	92.8	90.5	86.5	77.8	95.8	87.5
DINOv2											
DINOv2-S/14	53.8	50.8	69.1	52.8	71.8	53.4	52.4	62.6	45.3	51.7	56.4
DINOv2-B/14	59.5	54.5	66.6	49.7	67.9	55.4	55.6	60.2	44.1	57.6	57.1
DINOv2-L/14	61.1	60.6	68.5	52.5	68.6	61.9	58.4	63.7	57.3	59.6	61.2
DINOv2-H/14	62.3	58.1	69.4	55.1	66.3	60.8	59.8	65.1	59.6	58.1	61.4

- **Feature Extraction:** One forward pass over the 2,000-image reference set.
- **Linear Probe Training:** Optimization of 8,194 parameters in a single FC layer.

Tab. 7 compares the computational requirements of our method with prior detectors. The baselines differ substantially in both data scale and training cost: CNNSpot and UniverFD require optimization over large-scale datasets (~720k images), while CoDE and Community-Forensics introduce even heavier end-to-end training pipelines (5–12M images and tens of millions of trainable parameters). AIDE uses a ConvNeXt-based OpenCLIP backbone trained on 8×A100 GPUs and remains considerably more expensive than any frozen-feature approach.

In contrast, our method optimizes only a single linear layer with 8,194 parameters, and its dominant computational cost is a one-time feature extraction pass over the 2k reference images (approximately 3 minutes on a single RTX 5090). The training free variant removes this step entirely by performing no parameter updates. Consequently, our linear probe reduces the training cost by over **300×** relative to CNNSpot, while maintaining competitive performance across all benchmarks.

B. Impact of Model Scaling and Architecture

In this section, we systematically examine how model scale and architectural design affect cross-generator generalization. We evaluate a broad range of capacities within the DINO family, including **DINOv3: ViT-S/B/L/H-16** [46] and **DINOv2: ViT-S/B/L/H-14** [37], to assess whether larger backbones consistently yield stronger robustness under our evaluation protocol.

These two model families offer a natural architectural comparison: DINOv3 adopts a more recent self-supervised training strategy and scaling recipe, while DINOv2 represents an earlier generation of feature learning. Evaluating both across multiple scales allows us to analyze how much cross-generator performance is driven by model capacity versus architectural and training differences.

All experiments in this section follow a unified evaluation protocol: a lightweight linear probe is trained on the LDM reference set (1k real and 1k fake), and performance is evaluated on the So-Fake-OOD benchmark using the proposed FGTS representation.

The results in Tab. 8 and Fig. 7 show that DINOv3 exhibits a clear and consistent scaling trend. Accuracy in-

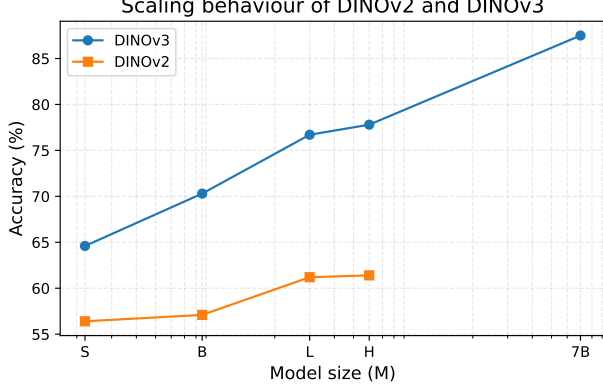


Figure 7. **Scaling behavior of DINOv2 and DINOv3 on cross-generator detection.** Accuracy on So-Fake-OOD is plotted against model size (millions of parameters). DINOv3 exhibits a clear scaling trend from ViT-S/16 to ViT-7B, while DINOv2 shows limited improvement with increased capacity.

creases steadily from ViT-S/16 to ViT-7B, indicating that both model capacity and the updated training strategy contribute to stronger cross-generator robustness. In comparison, DINOv2 shows only mild gains from S/14 to H/14, and its overall performance remains noticeably lower than DINOv3 at similar scales. This suggests that capacity alone is not sufficient without the architectural and training improvements present in DINOv3.

Given the smooth upward trajectory of DINOv3, particularly the substantial improvement at the 7B scale, it is reasonable to expect that further increasing **model capacity**, together with training strategies that reinforce global consistency, may continue to enhance cross-generator generalization.

C. Effectiveness of Fisher-Guided Token Selection (FGTS)

In this section, we evaluate the effectiveness of the proposed FGTS method under training free setting. We compare four DINOv3 models across three inference configurations: (1) using all tokens, (2) using only patch tokens, and (3) using the FGTS-selected token subset. This comparison allows us to isolate the contribution of token selection and to assess how much discriminative information FGTS preserves relative to full token configuration.

As illustrated in Fig. 8, FGTS consistently outperforms both the patch-only baseline and the all-token approach across all DINOv3 scales. Notably, FGTS achieves substantial improvements over using all tokens: +2.6% on ViT-B, +3.6% on ViT-L, and a striking +6.1% on ViT-H. The performance gap narrows to +1.6% on ViT-7B, suggesting that the largest models encode authenticity cues with sufficient redundancy that even naive token aggregation becomes ef-

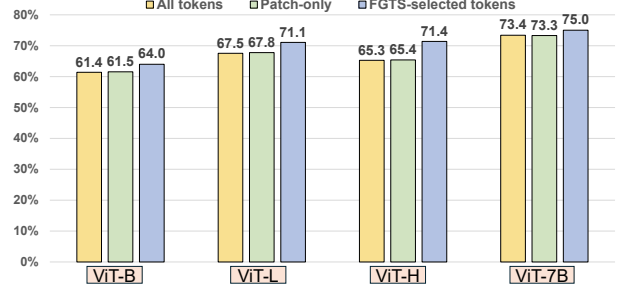


Figure 8. **Effectiveness of FGTS across DINOv3 scales.** FGTS consistently outperforms the patch-only baseline and approaches all-token accuracy across ViT-B/16 to ViT-7B.

fective. For mid-scale models, however, FGTS plays a critical role in filtering out noisy or less discriminative tokens that would otherwise dilute the authenticity signal. These results demonstrate that FGTS provides a more effective representation than naive token aggregation, with particularly significant benefits at mid-to-large scales.

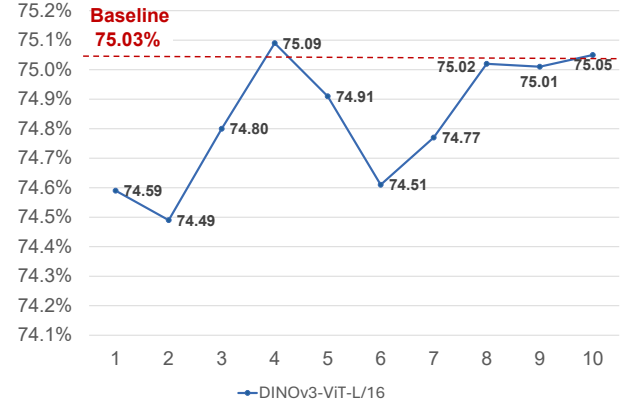


Figure 9. **Effect of large-scale fine-tuning on DINOv3-L/16.** Accuracy on So-Fake-OOD across 10 epochs fluctuates around the frozen baseline, indicating that supervised adaptation fails to strengthen cross-generator representations.

D. Analysis of Large-Scale Adaptation

To further understand the limitations of large-scale adaptation, we analyze how heavy supervised fine-tuning affects the intrinsic representation quality of DINOv3. We examine whether large-scale fine-tuning on diverse training categories can enhance cross-generator generalization. To investigate this systematically, we follow the data construction protocol of UniverFD [36], which uses 20 semantic categories such as *airplane*, *car*, and *dog*. For each category, we collect the corresponding real images and generate an equal number of synthetic images using LDM [42]. This mirrors the traditional deepfake training setup employed by

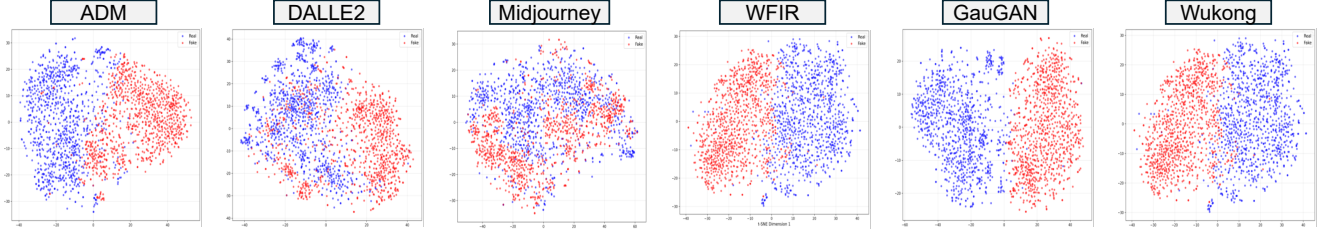


Figure 10. **t-SNE visualization of feature distributions** (Blue: Real, Red: Fake). The **first three** generators (ADM, DALLE-2, Midjourney) exhibit significant feature entanglement, which explains the performance gap compared to fine-tuned methods. In contrast, the **last three** (WFIR, GauGAN, Wukong) show clear linear separability, validating the effectiveness of our frozen backbone in capturing structural artifacts.

prior work, except that we use LDM rather than ProGAN to generate synthetic images. In total, the fine-tuning dataset contains 360k real images and 360k LDM generated images.

The model is fine-tuned for 10 epochs, and all other training configurations follow the settings described in Section A.1. We evaluate the resulting representations using a linear probe on top of DINOv3-L/16 and report performance on the So-Fake-OOD benchmark. This setup directly measures how large-scale supervised adaptation affects the model’s ability to retain global authenticity cues.

As shown in Fig. 9, large-scale supervised fine-tuning fails to improve the cross-generator generalization of DINOv3-L/16. Across 10 epochs of training, accuracy on So-Fake-OOD fluctuates around the frozen baseline (75.03%), ranging from 74.49% to 75.09%. The best checkpoint at epoch 4 achieves 75.09%, showing only marginal improvement over the baseline, while several epochs exhibit slight performance drops, indicating instability during adaptation.

This suggests that the global authenticity cues encoded during DINOv3 pretraining are already highly effective and difficult to enhance through standard supervised learning. The lack of consistent improvement, combined with training instability, indicates that large-scale fine-tuning offers little benefit for cross-generator generalization in this setting. In contrast, the lightweight linear probe achieves comparable performance without the computational cost and potential risks of full backbone adaptation.

E. Analysis of Performance Variations Across Generators

While FGTS demonstrates strong overall generalization, its performance varies across generators in AIGCDetectionBenchmark [61]. We achieve over 95% accuracy on certain generators (e.g., WFIR, VQDM) but observe reduced performance on Midjourney (77.32%) and DALLE-2 (77.00%). We investigate the underlying causes through feature space and frequency domain analysis.

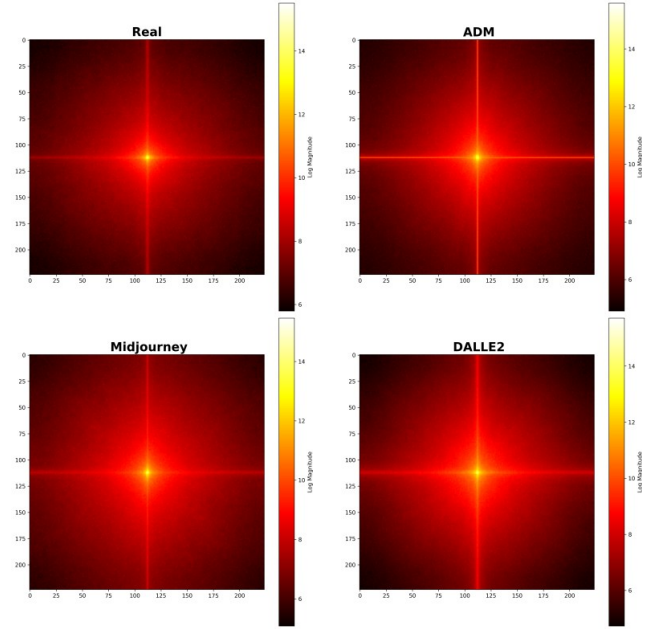


Figure 11. **Frequency spectrum analysis.** ADM, DALLE-2, and Midjourney exhibit frequency distributions nearly identical to real images, with energy concentrated in low-frequency components (central bright cross). This high-fidelity low-frequency replication explains the feature entanglement observed in Fig. 10 and the detection challenges for frozen low-frequency-biased representations.

t-SNE Visualization. Fig. 10 visualizes frozen DINOv3 features via t-SNE for six generators. The results reveal two distinct patterns: (1) **Challenging generators** (ADM, DALLE-2, Midjourney) exhibit significant real/fake feature entanglement, making linear separation difficult. (2) **Separable generators** (WFIR, GauGAN, Wukong) show clear linear separability with distinct real/fake clusters. This pattern directly correlates with their respective detection accuracy.

Frequency Domain Analysis. Fig. 11 shows frequency spectra for Real, ADM, Midjourney, and DALLE-2. All

four exhibit comparable low-frequency energy concentration (central cross pattern) with minimal high-frequency content. Notably, challenging generators (ADM, DALLE-2, Midjourney) show closer spectral similarity to real images than separable generators. This observation is consistent with our finding that DINOv3 relies predominantly on low-frequency global structure: when generators produce low-frequency patterns similar to real images, frozen features exhibit reduced discriminability.

Although FGTS achieves the best cross-dataset generalization overall, the above analysis reveals room for further improvement on these generators that successfully replicate low-frequency characteristics of real images.

References

- [1] Andrew P. Aitken, Christian Ledig, Lucas Theis, Jose Caballero, Zehan Wang, and Wenzhe Shi. Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. *Arxiv*, 2017. 3
- [2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: BERT pre-training of image transformers. In *ICLR*, 2022. 9
- [3] Xiuli Bi, Bo Liu, Fan Yang, Bin Xiao, Weisheng Li, Gao Huang, and Pamela C. Cosman. Detecting generated images by real images only. *Arxiv*, 2023. 7, 8
- [4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3
- [5] Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. Contrasting deep-fakes diffusion via contrastive learning and global-local similarities. In *ECCV*, 2024. 1, 7, 8, 9, 10
- [6] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP*, 2023. 3
- [7] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with CLIP. In *CVPR*, 2024. 1
- [8] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024. 3
- [9] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 1
- [10] 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. In *ICLR*, 2021. 2
- [11] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *CVPR*, 2020. 3, 8
- [12] Tarik Dzanic, Karan Shah, and Freddie D. Witherden. Fourier spectrum discrepancies in deep network generated images. In *NeurIPS*, 2020. 3, 8
- [13] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 1936. 2
- [14] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, 2020. 3, 8
- [15] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. *Arxiv*, 2020. 7, 8
- [16] Quanquan Gu, Zhenhui Li, and Jiawei Han. Generalized fisher score for feature selection. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 2012. 2
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7, 9
- [18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 9
- [19] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Arxiv*, 2021. 1
- [20] Yuqi Hu, Longguang Wang, Xian Liu, Ling-Hao Chen, Yuwei Guo, Yukai Shi, Ce Liu, Anyi Rao, Zeyu Wang, and Hui Xiong. Simulating the real world: A unified survey of multimodal generative models. *Arxiv*, 2025. 1
- [21] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. SIDA: social media image deepfake detection, localization and explanation with large multimodal model. *CVPR*, 2025. 8
- [22] Zhenglin Huang, Tianxiao Li, Xiangtai Li, Haiquan Wen, Yiwei He, Jiangning Zhang, Hao Fei, Xi Yang, Xiaowei Huang, Bei Peng, and Guangliang Cheng. So-fake: Benchmarking and explaining social media image forgery detection. *Arxiv*, 2025. 1, 3, 7
- [23] Nick Jiang, Amil Dravid, Alexei A. Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. *Arxiv*, 2025. 3
- [24] Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *ICIP*, 2022. 3, 7, 8
- [25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018. 7, 9
- [26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 3
- [27] Jiawei Li, Fanrui Zhang, Jiaying Zhu, Esther Sun, Qiang Zhang, and Zheng-Jun Zha. Forgerypt: Multimodal large language model for explainable image forgery detection and localization. *Arxiv*, 2024. 8

- [28] Yixuan Li, Yu Tian, Yipo Huang, Wei Lu, Shiqi Wang, Weisi Lin, and Anderson Rocha. Fakescope: Large multimodal expert model for transparent ai-generated image forensics. *Arxiv*, 2025. 8
- [29] Kaiqing Lin, Yuzhen Lin, Weixiang Li, Taiping Yao, and Bin Li. Standing on the shoulders of giants: Reprogramming visual-language model for general deepfake detection. In *AAAI*, 2025. 1
- [30] Zhengzhe Liu, Xiaojuan Qi, and Philip H. S. Torr. Global texture enhancement for fake face detection in the wild. In *CVPR*, 2020. 7
- [31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 7
- [32] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 9
- [33] Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. SLIP: self-supervision meets language-image pre-training. In *ECCV*, 2022. 9
- [34] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, B. S. Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H. Bappy, and Amit K. Roy-Chowdhury. Detecting GAN generated fake images using co-occurrence matrices. In *Electronic imaging*, 2019. 8
- [35] Sophie J. Nightingale and Hany Farid. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *PNAS*, 2022. 3
- [36] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023. 1, 7, 8, 9, 10, 11
- [37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. 3, 9, 10
- [38] Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *CVPR*, 2025. 7, 9, 10
- [39] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020. 7
- [40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 7, 9
- [41] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, 2021. 6
- [42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 9, 11
- [43] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 3
- [44] Michael S. Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: What can 8 learned tokens do for images and videos? *Arxiv*, 2021. 6
- [45] Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *Arxiv*, 2024. 1
- [46] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *Arxiv*, 2025. 2, 3, 9, 10
- [47] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 9
- [48] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *CVPR*, 2023. 7, 8
- [49] Chuangchuang Tan, Huan Liu, 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 *CVPR*, 2024. 7, 8
- [50] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 7
- [51] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, 2020. 7, 8, 9, 10
- [52] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *ICCV*, 2023. 7, 8
- [53] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan L. Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022. 9
- [54] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 9
- [55] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *AAAI*, 2022. 6

- [56] Zhipei Xu, Xuanyu Zhang, Runyi Li, Zecheng Tang, Qing Huang, and Jian Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *Arxiv*, 2024. [8](#)
- [57] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *ICLR*, 2025. [1](#), [7](#), [8](#), [9](#), [10](#)
- [58] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *Arxiv*, 2015. [9](#)
- [59] Ning Yu, Larry Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing GAN fingerprints. In *ICCV*, 2019. [8](#)
- [60] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. *Arxiv*, 2019. [7](#)
- [61] Nan Zhong, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. *Arxiv*, 2023. [7](#), [8](#), [12](#)
- [62] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan L. Yuille, and Tao Kong. ibot: Image BERT pre-training with online tokenizer. *Arxiv*, 2021. [9](#)
- [63] 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 *NeurIPS*, 2023. [2](#), [7](#)