DPG-PROBE: RESOLVING THE SEMANTIC-ARTIFACT DILEMMA VIA DUAL-PRESERVING GRADIENTS FOR AIGI FORENSICS

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid evolution of Artificial Intelligence Generated Imagery presents severe forensic challenges characterized by two primary hurdles: the unprecedented fidelity and diversity of synthetic content, and the critical demand for robust, 'in-the-wild' real-world AIGI detection. To tackle these, current methods are largely based on advanced vision foundation models because they have already observed a vast visual world. However, this is precisely the fundamental reason why detection struggles to generalize. In the CLS space, real and fake images under the same paradigm theme are extremely close, meaning that when your test semantics deviate from the training set, the boundary structure moves with the object, rather than the intrinsic artifact.

Addressing this limitation requires treating AIGI detection as a matter of provenance rather than standard semantic classification. Let f represent a frozen VFM and T a semantic-content and signal-structure preserving transformation (i.e., orthogonal rotations). While $f(x)$ and $f(T(x))$ should match on the core content of the image, any residual discrepancies are captured by hidden gradient directions ($g(x)$) sensitive to T. Therefore, we propose DPG-Probe, which leverages the residual discrepancies between $f(x)$ and $f(T(x))$ as a training-free, label-free probe. By taking a single backward pass of a self-consistency loss on intermediate CLS tokens through the frozen VFM, the resulting CLS gradients ($g(x)$) capture a structural common pattern of synthetic image formation rather than surface content. Because the gradient probe relies on dual-preserving transformations, it suppresses high-level semantic shifts and maintains superior stability under wild degradation compared to pixel-level artifacts. Comprehensive evaluations demonstrate that DPG-Probe significantly enhances cross-domain generalization and real-world robustness over prior detectors.

1 INTRODUCTION

AI-generated images now match the visual quality and landscape of photographs: identical objects, matching scenes, and frequently, the same aesthetic styles. Consequently, a forensic detector that asks what is depicted in the image is addressing the wrong question. Forensic analysis must instead determine how the image was produced, whether its pixels originated from a physical camera scene or an AI generator and decoder. This distinction is the essence of provenance, not a mere semantic category.

Frozen vision foundation models (VFMs) present an appealing substrate for this task because they have already internalized a vast visual world without requiring supervision from real/fake labels He et al. (2026); Zhou et al. (2026b); Ojha et al. (2023); Lin et al. (2024); Ding et al. (2025); Yan et al. (2025a; 2024). Yet, standard usage patterns fall right into this trap. The forward $[\mathrm{CLS}]$ token acts as a content summary, and it is optimized to remain stable across crops, lighting shifts, and minor appearance changes while varying across object and scene transitions. As a result, real and fake images of identical subjects often cluster closely together in the $[\mathrm{CLS}]$ space, whereas two real images of different subjects are cast far apart. Classifiers or nearest-neighbor rules built directly on $[\mathrm{CLS}]$ inherit this geometry. Once the test semantics depart from the reference set, the resulting boundary structures shift according to the objects rather than the source.

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.

Conversely, the alternative approach is to hunt for generator-specific artifacts directly in the pixel space: frequency peaks, upsampling grids, or autoencoder residuals Chen et al. (2025b); Zhou et al. (2026a); Li et al. (2025a). While these traces are genuine, they reside at a scale easily destroyed by social media pipelines. Operations like resizing, re-compression, and blurring do not invalidate the synthetic nature of an image; rather, they erase the specific low-level patterns that a given detector was tuned to capture.

To bridge these two extremes, we define a new function to extract visual representations. Let f represent a frozen VFM, and x an input image. Instead of static representations $f(x)$, we capture how f responds when x undergoes a transformation that ought to leave its content and signal features invariant. Let T represent the transformation that preserves objects (i.e., preserves semantic content and signal-structure information) on the canvas. T changes some dimensions of the intermediate [CLS] and leaves others nearly fixed. The residual $f_\ell(x) - f_\ell(T(x))$ lives in the dimensions the still move with T. Cameras and generators encode fundamentally distinct signatures within these dimensions. The first-order manifestation of this disagreement is formalized as:

$$g(x) = \nabla_h \left\| f_\ell(x) - f_\ell(T(x)) \right\|^2 \qquad (1)$$

where h denotes an intermediate [CLS] state, and ℓ indexes a middle layer.

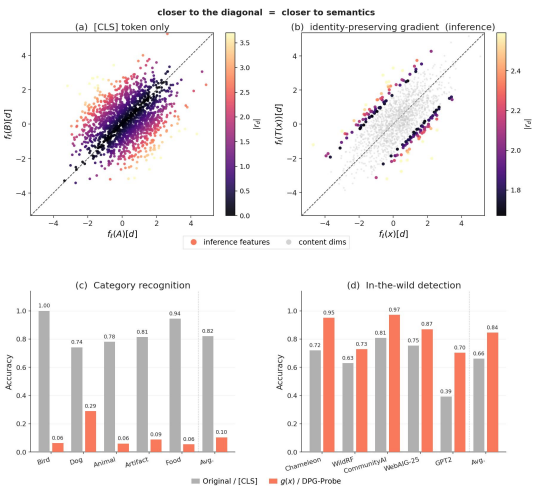


Figure 1: Dual-preserving geometry and its downstream effect. (a–b) Static [CLS] vs. $g(x)$. (c–d) Original / [CLS] preserve ImageNet semantics but $g(x)$ with DPG-Probe improves in-the-wild detection.

As illustrated in Figure 1, this geometric transformation reveals the core intuition behind our approach: while the static [CLS] token is heavily mired in content and semantic mixtures, our dual-preserving transformation T pins the shared semantic content dimensions (gray) rigidly along the diagonal. This cleanly isolates and lifts up the generalizable forensic inference features (colored). This isolation is reflected downstream: semantic category scores drop under $g(x)$ while in-the-wild detection improves with DPG-Probe (Fig. 1, bottom). By replacing static [CLS] activations with $g(x)$ as a dynamic forensic probe, we bypass the semantic traps of standard classification and directly capture the common pattern cues that remain consistent across diverse synthetic images.

We operationalize this probing mechanism as DPG-Probe. The encoder remains strictly frozen as MetaCLIP-2. Each image is transformed into a small set of dual-preserving views; a self-consistency loss is evaluated on intermediate [CLS] tokens; a single backward pass yields $g(x)$; and final detection is performed via nearest-neighbor retrieval against a labeled reference bank. The method involves zero training on forensic tasks. Consequently, its validity rests entirely on the hypothesis that provenance resides within T-sensitivity rather than semantic positioning. The remainder of this paper puts this hypothesis to the rigorous test.

2 Theoretical Insights and Design Rationale

Three foundational consequences follow from the formulation 1, dictating the design of our method rather than serving as post-hoc rationalizations of empirical results. Fig. 2 visualizes the Layer-32 mixed-pool PCA projections under four distinct formulations. The top row is colored by real/fake provenance, and the bottom row by underlying datasets, revealing how each representation organizes the visual space.

The Semantic Trap of [CLS] Tokens (Column 1). The standard [CLS] feature space organizes images strictly by dataset and semantic content rather than provenance. As shown in the bottom-left panel, distinct dataset clusters dominate the geometry, while real (blue) and fake (red) samples remain



Figure 2: PCA visualization of Layer-32 feature spaces under four distinct formulations, with the top row colored by real/fake provenance and the bottom row colored by underlying datasets.

hopelessly intertwined within each cluster. This explains why a baseline using $[\mathrm{CLS}]$ remains fragile under semantic shifts, even when keeping the identical frozen backbone and reference bank.

The Collapse of Scalar Consistency (Column 2). $g(x)$ is not merely the scalar magnitude $\|f(x) - f(T(x))\|$. Discarding directional information by reducing the transformation residual to a scalar magnitude ($\|f(x) - f(T(x))\|$) completely destroys the feature's discriminative geometry. The projection collapses into an undifferentiated, isotropic blob where real and fake distributions completely overlap. What a k-nearest-neighbor rule requires is the direction of the residual, rather than a collapsed 1-D scalar score.

The Contamination of Content-Breaking Views (Column 3). Employing spatial partitioning or disjoint crops (denoted as PatchQuad grad) introduces severe content noise. Without strict semantic-content preservation, the gradient is contaminated by structural mismatches (e.g., swapping foreground objects or background regions). Consequently, the feature space distorts into an erratic, elongated streak that neither isolates the source nor preserves semantic stability. Common signal-corrupting augmentations (e.g., Gaussian blur or JPEG compression) introduce irrecoverable signal loss in high-frequency regimes, whereas identity-breaking spatial splits inject cross-region semantic noise. Both degrade $g(x)$ compared to purely orthogonal rotations.

The Clean Separation of Invariance-Failure Gradients (Column 4). Our formulation leveraging rotation-induced, dual-preserving (semantic-content and signal-structure) gradients $g(x)$ uniquely resolves these limitations. As shown in the rightmost column, real and fake samples form well-segregated clusters, while the multi-source dataset colors (bottom row) harmoniously converge. By probing how internal hidden dimensions must shift to restore transformation invariance, the representation successfully strips away semantic interference while retaining robust, cross-domain forensic signatures.

3 Shared Structure in $g(x)$, Not in $f_\ell(x)$

Sec. 2 showed that rotation gradients organize a mixed pool by provenance rather than by dataset. A natural follow-up is whether dual-preserving gradients $g(x)$ exhibit shared cross-dimensional structure among synthetic images from distinct generators.

Fig. 3 answers this at the micro level by focusing exclusively on fake groups from four diverse datasets (Now-sem, Holmes, Chameleon, WildRF). For each representation ($[\mathrm{CLS}]$ vs. $g(x)$), we select the top-48 layer-32 dimensions with the largest real/fake Fisher separation ratio, compute a per-dataset

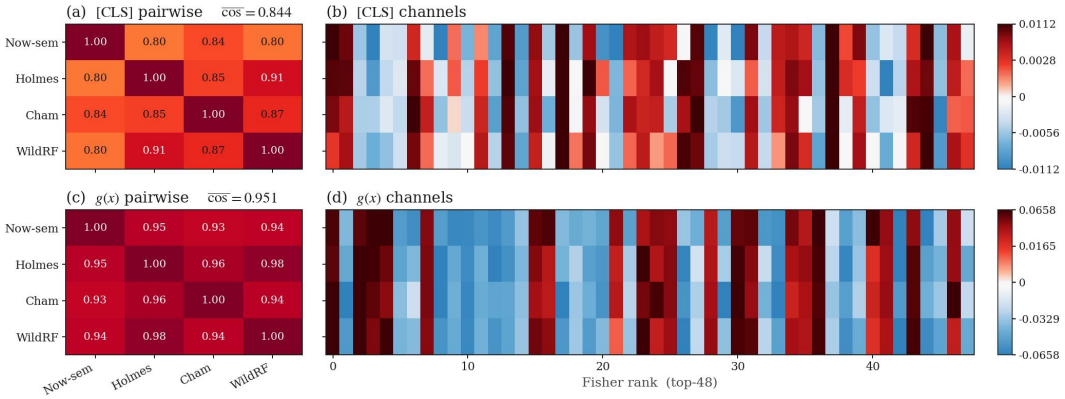


Figure 3: Fake-only micro-profiles at layer 32: [CLS] (top) vs. dual-preserving gradients $g(x)$ (bottom). Left: 4×4 pairwise cosine among dataset-level fake means in top-48 Fisher dimensions ($\overline{\cos}$=0.844 vs. 0.951). Right: corresponding channel heatmaps (top-48 Fisher ranks).

fake group-mean profile in this subspace, and report two complementary views. Panels (a) and (c) summarize cross-dataset agreement via pairwise cosine similarity among the four fake signatures (off-diagonal mean $\overline{\cos}$). Panels (b) and (d) plot the same signatures as heatmaps, with Fisher rank on the horizontal axis and color indicating signed normalized activation.

Under forward [CLS], fake signatures remain only moderately aligned across generators ($\overline{\cos}$=0.844; Fig. 3a), and the channel map is largely unstructured and dataset-specific (Fig. 3b), consistent with semantics and scene content dominating static activations. In contrast, $g(x)$ yields substantially tighter cross-dataset concordance ($\overline{\cos}$=0.951; Fig. 3c) and recurrent high-magnitude columns across all four benchmarks (synchronized vertical bands in Fig. 3d). This provides direct evidence that synthetic images from heterogeneous sources share a directional coordinate system in $g(x)$ after dual-preserving rotation, which nearest-neighbor retrieval can exploit to bypass semantic interference.

4 Methodology

4.1 Overview of the DPG-Probe Framework

Given an arbitrary input image x, our goal is to map it into a generic forensic representation $\mathcal{F}(x) \in \mathbb{R}^D$ without updating the parameters Θ of a pre-trained Vision Foundation Model (VFM). As illustrated in Figure 4, the proposed **DPG-Probe** framework operates in a non-parametric, training-free strategy. It consists of two seamlessly integrated components:

- **Dual-Preserving Gradient Extraction Module (ϕ_{grad}):** A uniform feature transformation module that probes intermediate VFM representations via dual-preserving spatial rotations, distilling a semantic-purified gradient descriptor $\hat{f}(x)$ for both reference and target images.
- **Non-Parametric Metric Retrieval:** A training-free evaluation protocol that performs zero-shot cross-target forgery detection via k-Nearest Neighbor (kNN) retrieval in a normalized gradient manifold.

4.2 Dual-Preserving Gradient Extraction Module

The core of our approach lies in constructing a gradient-based representation extraction operator $\phi_{\text{grad}}(\cdot)$. For any given image x (whether from a reference bank or an unobserved target domain), $\phi_{\text{grad}}(x)$ extracts its forensic signature via the following multi-view consistency perturbation and backward probe.

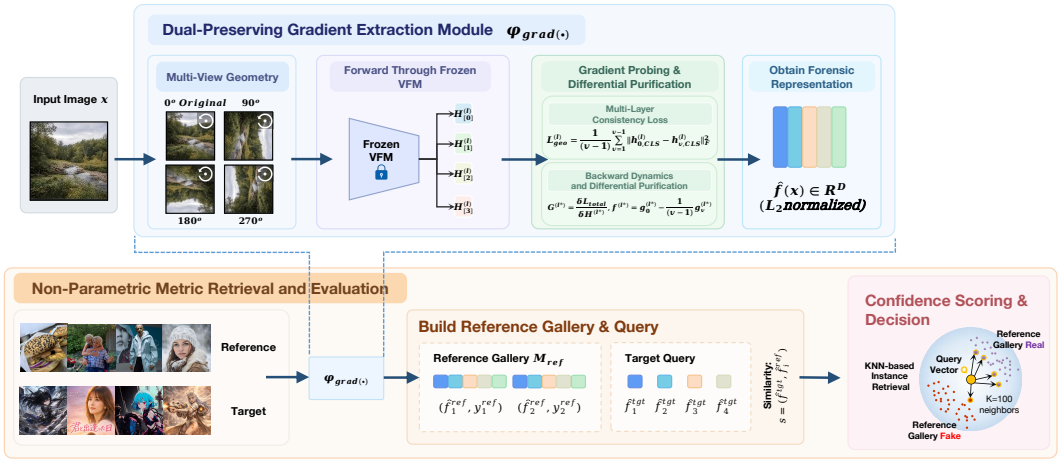


Figure 4: Overview of the DPG-Probe framework compared with existing paradigms and its performance evaluation.

4.2.1 Geometry-Mode Multi-View Transformation

Formally, given an input anchor image x, we construct a multi-view perturbation set $\mathcal{X} = \{\tau_0(x), \tau_1(x), \ldots, \tau_{V-1}(x)\}$, where V denotes the total number of augmented views and $\tau_0(x) = x$ represents the pristine identity view. To guarantee that the transformation preserves high-level semantic topology while exposing intrinsic generative artifacts, we adopt orthogonal spatial rotations:

$$\tau_v(x) = \text{Rotate}(x, \theta_v), \quad \theta_v \in \{0°, 90°, 180°, 270°\} \tag{2}$$

where $V = 4$. The batched representation $\mathcal{X} \in \mathbb{R}^{V \times C \times H \times W}$ is formed by concatenating these views along the batch dimension.

4.2.2 Multi-View Multi-Layer Aggregated Gradient Probing

The batched tensor $\mathcal{X}$ is forwarded through a frozen VFM $\Phi(\cdot; \Theta)$. To capture rich, multi-scale consistency constraints without spatial token interference, we extract intermediate hidden states at a subset of layers $\mathcal{L} = \{32, 36, 40, 44, 48\}$. Let $\mathcal{H}^{(l)} \in \mathbb{R}^{V \times N \times D}$ denote the hidden state tensor at layer l, where N is the token sequence length ([CLS] token plus spatial patch tokens) and D is the feature dimension. The global [CLS] token representation for view v is specifically denoted as $h^{(l)}_{v,\text{CLS}} \in \mathbb{R}^D$.

Multi-Layer Unsupervised Consistency Objective. At each sampled layer $l \in \mathcal{L}$, we compute a geometry-mode consistency loss $\mathcal{L}^{(l)}_{\text{geo}}$ applied exclusively to the [CLS] token. This penalizes the discrepancy between the pristine anchor ($v = 0$) and its perturbed counterparts ($v \in \{1, \ldots, V-1\}$):

$$\mathcal{L}^{(l)}_{\text{geo}} = \frac{1}{V-1} \sum_{v=1}^{V-1} \text{MSE}\left(h^{(l)}_{0,\text{CLS}}, h^{(l)}_{v,\text{CLS}}\right) \tag{3}$$

To form a unified backward dynamic that captures multi-scale semantic and structural inconsistencies simultaneously, the total consistency loss $\mathcal{L}_{\text{total}}$ is averaged across all sampled layers:

$$\mathcal{L}_{\text{total}} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \mathcal{L}^{(l)}_{\text{geo}} \tag{4}$$

Backward Dynamics and Differential Purification. By executing a single backward pass of $\mathcal{L}_{\text{total}}$, we compute the gradient tensor with respect to the full hidden states at a target layer l^* (e.g., $l^* = 32$):

$$\mathcal{G}^{(l^*)} = \frac{\partial \mathcal{L}_{\text{total}}}{\partial \mathcal{H}^{(l^*)}} \in \mathbb{R}^{V \times N \times D} \tag{5}$$

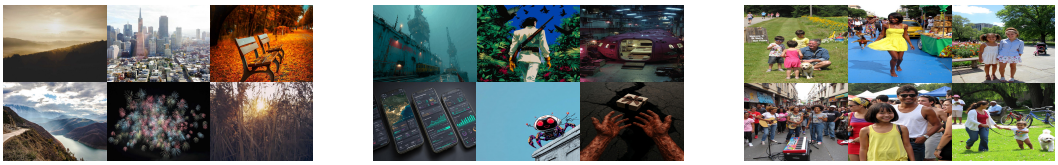


Figure 5: Visual samples from the WebAIG-25 dataset

To isolate the optimal forensic signature and strip away residual semantic shortcuts, we extract the gradient vector corresponding strictly to the [CLS] token and compute the differential residual:

$$g_v^{(l^*)} = \mathcal{G}^{(l^*)}[v, \text{CLS}, :], \quad v \in \{0, 1, \ldots, V-1\} \tag{6}$$

$$f^{(l^*)} = g_0^{(l^*)} - \frac{1}{V-1} \sum_{v=1}^{V-1} g_v^{(l^*)} \tag{7}$$

The final normalized forensic representation $\hat{f}(x) = \phi_{\text{grad}}(x)$ is derived via ℓ_2-normalization: $\hat{f}(x) = f^{(l^*)} / \|f^{(l^*)}\|_2$.

4.3 Non-Parametric Forensic Metric Retrieval

Once images are mapped into the purified gradient space via $\phi_{\text{grad}}(\cdot)$, cross-target AIGI detection is performed via non-parametric metric retrieval, bypassing the need for trainable classification heads ($e.g.$, MLPs) that suffer from semantic overfitting.

Reference Index Construction. We build a fixed reference gallery $\mathcal{M}_{\text{ref}}$ from the full training split of a known source dataset $\mathcal{D}_{\text{ref}} = \{(x_i^{\text{ref}}, y_i^{\text{ref}})\}_{i=1}^{M_{\text{ref}}}$ ($e.g.$, GenImage Stable Diffusion v1.4 train split), where $y_i^{\text{ref}} \in \{0, 1\}$ (0 for real, 1 for fake). Each reference image is passed through $\phi_{\text{grad}}(\cdot)$ to construct the gallery:

$$\mathcal{M}_{\text{ref}} = \left\{ \left(\phi_{\text{grad}}(x_i^{\text{ref}}),\, y_i^{\text{ref}} \right) \right\}_{i=1}^{M_{\text{ref}}} = \left\{ \left(\hat{f}_i^{\text{ref}},\, y_i^{\text{ref}} \right) \right\}_{i=1}^{M_{\text{ref}}} \tag{8}$$

Target Projection and kNN Confidence Scoring. Given an unobserved target query image $x_j^{\text{tgt}} \in \mathcal{D}_{\text{tgt}}$, we extract its descriptor $\hat{f}_j^{\text{tgt}} = \phi_{\text{grad}}(x_j^{\text{tgt}})$. Because descriptors are ℓ_2-normalized, similarity is evaluated via dot product:

$$\text{Sim}\left(\hat{f}_j^{\text{tgt}},\, \hat{f}_i^{\text{ref}} \right) = \langle \hat{f}_j^{\text{tgt}},\, \hat{f}_i^{\text{ref}} \rangle \tag{9}$$

Let $\Omega_k\left(\hat{f}_j^{\text{tgt}}\right)$ denote the top-k nearest neighbors retrieved from $\mathcal{M}_{\text{ref}}$. The final synthetic confidence score $s_j \in [0, 1]$ and binary decision $\hat{y}_j^{\text{tgt}}$ are defined as:

$$s_j = \frac{1}{k} \sum_{i \in \Omega_k\left(\hat{f}_j^{\text{tgt}}\right)} y_i^{\text{ref}}, \quad \hat{y}_j^{\text{tgt}} = \mathbb{I}(s_j > 0.5) \tag{10}$$

5 Experiments

5.1 Experimental Setup

5.1.1 Evaluation Datasets and Protocols

To rigorously assess the in-the-wild and semantic out-of-domain generalization of DPG-Probe, we organize our evaluation into two progressively challenging benchmark categories:

1. **Semantic OOD Benchmarks:** We evaluate on standard AIGI forensics benchmark **GenImage** Zhu et al. (2023) alongside its latest iteration **GenImage++**$_{v2}$ Zhou et al. (2026a). In this version, we expanded the dataset by introducing a highly deceptive and challenging newly released subset, *flux_amateur*. Additionally, we evaluate on the cutting-edge benchmarks **AIGIHolmes** Zhou et al. (2025), **AIGI-Now** Chen et al. (2025a), **AnimeDL-2M** Zhu et al. (2025) and **T2I-COREBENCH** Li et al. (2025b).
2. **In-the-Wild Datasets:** We benchmark on several unconstrained media datasets, including **Chameleon** Yan et al. (2025b), **WildRF** Cavia et al. (2024), **GPT-Image-2** Zewde et al. (2026) and **CommunityAI** Li et al. (2026).

   Furthermore, we introduce a custom-constructed, stringent blind dataset named **WebAIG-25** to serve as a decisive containment test against pre-training data leakage.

   - *Synthetic Split (midjourney_images):* Consists of AI-generated artworks scraped from *r/midjourney* posted strictly after the year 2025, guaranteed to succeed the training cut-off dates of all mainstream VFMs.
   - *Real Split (picsum_images):* Comprise high-resolution photographs of natural scenes.

5.1.2 Implementation Details

We adopt the frozen vision backbone of MetaCLIP2 Chuang et al. (2026) as our default Vision Foundation Model $\Phi(\cdot;\Theta)$ without optimizing any parameters. The neighborhood parameter k for the non-parametric metric retrieval is fixed to 100. All visual inputs are normalized and resized to 224×224 pixels following the VFM processor standards.

5.1.3 Compared Baselines

We comprehensively compare our DPG-Probe against two major schools of state-of-the-art (SOTA) AIGI forensic paradigms:

- **Semantic-Dominant Projections:** *CNNSpot* Wang et al. (2020), *FreqNet* Tan et al. (2024a), *Gram-Net* Liu et al. (2020), *UnivFD* Ojha et al. (2023), *LaDeDa* Cavia et al. (2024), *Effort* Yan et al. (2024), *AIDE* Yan et al. (2025b), *DGS-Net* Yan et al. (2025a).
- **Micro-Artifact Anomalies:** *NPR* Tan et al. (2024b), *SAFE* Li et al. (2025a), *DDA* Chen et al. (2025b), and *OMAT* Zhou et al. (2026a).

All compared methods utilize their officially released weights and maintain the exact model configurations as described in their original papers, without any additional local training.

5.2 Main Experiment I: Comparison with State-of-the-Art

5.2.1 Evaluation on In-the-Wild Benchmarks

Table 1: Performance on In-the-Wild Benchmarks.

Method	Chameleon			WildRF			CommunityAI			WebAIG-25			GPT2	Avg.
	Real	Fake	Avg.	Real	Fake	Avg.	Real	Fake	Avg.	Real	Fake	Avg.		
Semantic-Dominant Projections														
CNNSpot	0.979	0.128	0.554	0.959	0.290	0.625	0.969	0.112	0.541	0.971	0.083	0.527	0.019	0.453
FreqNet	0.985	0.090	0.538	0.731	0.559	0.645	0.977	0.128	0.553	0.670	0.260	0.465	0.232	0.487
Gram-Net	0.992	0.044	0.518	0.947	0.205	0.576	0.985	0.061	0.523	0.989	0.054	0.522	0.026	0.433
UnivFD	0.763	0.441	0.602	0.693	0.637	0.665	0.778	0.495	0.636	0.592	0.298	0.445	0.442	0.558
LaDeDa	0.994	0.015	0.504	**0.988**	0.119	0.554	0.986	0.026	0.506	**0.998**	0.110	0.554	0.026	0.429
Effort	0.394	0.782	0.588	0.179	0.955	0.567	0.225	0.840	0.533	0.154	0.905	0.530	**0.797**	0.603
AIDE	0.944	0.203	0.574	0.973	0.195	0.584	0.990	0.093	0.542	0.303	**0.971**	0.637	0.151	0.498
DGS-Net	0.997	0.021	0.578	0.937	0.363	0.633	**0.998**	0.030	0.514	**0.998**	0.251	0.625	0.031	0.476
Micro-Artifact Anomalies														
NPR	**0.999**	0.046	0.523	0.980	0.243	0.612	**0.998**	0.076	0.537	0.935	0.349	0.642	0.039	0.471
SAFE	0.993	0.046	0.520	0.918	0.309	0.613	0.984	0.080	0.532	0.997	0.213	0.605	0.105	0.475
DDA	0.940	0.708	0.824	0.899	0.908	**0.904**	0.968	0.725	0.847	0.655	0.931	0.793	0.203	0.714
OMAT	0.899	0.359	0.629	0.633	0.715	0.674	0.853	0.414	0.634	0.876	0.471	0.674	0.154	0.553
Ours	0.962	**0.937**	**0.950**	0.475	**0.983**	0.729	0.965	**0.978**	**0.971**	0.799	0.942	**0.870**	0.704	**0.845**

To evaluate the absolute limits of in-the-wild forensic generalizability, we compare DPG-Probe against major SOTA paradigms on five unconstrained internet-scraped datasets. As quantified in Table 1, standard SOTA paradigms suffer from severe performance degradation under unconstrained online environments, whereas our training-free framework establishes a new state-of-the-art benchmark with an average Accuracy of 0.845. This validates our hypothesis that micro-artifact detectors ($e.g.$, SAFE, OMAT) fail drastically, as transient pixel-level and frequency-level physical fingerprints are effortlessly obliterated by realistic online post-processing during social network transmission.

Table 2: Performance on Out-of-Semantic-Domain Benchmarks.

Method	GenImage++	AnimeBench	CoreBench	GenImage			AIGIHolmes			AIGI-Now						Avg.
	Avg.	Avg.	Avg.	Real	Fake	Avg.	Real	Fake	Avg.	pix-Real	pix-Fake	pix-Avg.	sem-Real	sem-Fake	sem-Avg.	
Semantic-Dominant Projections																
CNNSpot	0.221	0.057	0.156	**0.996**	0.416	0.706	**0.996**	0.149	0.573	**0.997**	0.106	0.552	0.999	0.060	0.529	0.399
FreqNet	0.874	0.086	0.826	0.899	0.780	0.840	0.864	0.613	0.738	0.873	0.623	0.748	0.855	0.137	0.496	0.658
Gram-Net	0.327	0.017	0.253	0.983	0.528	0.756	0.980	0.231	0.606	0.980	0.194	0.587	**1.000**	0.007	0.504	0.435
UnivFD	0.820	0.782	0.764	0.939	0.695	0.817	0.939	0.628	0.784	0.938	0.574	0.756	0.969	0.110	0.539	0.752
LaDeDa	0.464	0.020	0.355	0.985	0.354	0.669	0.993	0.299	0.646	0.992	0.190	0.591	0.994	0.006	0.500	0.464
Effort	0.957	0.738	0.938	0.745	0.868	0.807	0.614	0.831	0.722	0.600	0.871	0.736	0.987	0.302	0.644	0.792
AIDE	0.930	0.066	0.776	0.919	0.616	0.768	0.891	0.611	0.751	0.880	0.584	0.732	0.971	0.014	0.493	0.645
DGS-Net	0.350	0.022	0.400	0.985	0.617	0.801	0.973	0.364	0.668	0.968	0.215	0.592	0.998	0.035	0.517	0.478
Micro-Artifact Anomalies																
NPR	0.890	0.060	0.778	0.990	0.685	0.838	0.991	0.596	0.794	0.994	0.481	0.737	**1.000**	0.001	0.500	0.657
SAFE	0.872	0.067	0.847	0.978	0.694	0.836	0.961	0.632	0.796	0.967	0.540	0.754	0.975	0.002	0.489	0.666
DDA	0.944	0.651	0.897	0.979	0.801	0.890	0.990	**0.937**	**0.963**	0.991	0.677	0.834	0.967	0.147	0.557	0.820
OMAT	0.931	0.240	0.840	0.948	**0.943**	**0.945**	0.944	0.810	0.877	0.944	0.558	0.751	0.893	0.067	0.480	0.723
Ours	**0.964**	**0.950**	**0.953**	0.979	0.792	0.886	0.975	**0.937**	0.956	0.972	**0.940**	**0.956**	0.944	**0.912**	**0.928**	**0.942**

5.2.2 Evaluation on Semantic Out-of-Domain Benchmarks

Table 2 reports the DPG-Probe zero-shot detection performance against completely unobserved distributions and wide semantic span synthesis models, where our framework demonstrates decisive superiority with an overall average Accuracy of 0.942. Crucially, when pushed into domains with severe domain style shifting, such as the artistic layouts in *AnimeBench*, standard baselines collapse comprehensively. In sharp contrast, DPG-Probe consistently bridges these vast distribution gaps, capturing 0.952 on AnimeBench, 0.948 on AIGIHolmes, and an unprecedented 0.945 / 0.929 across the pixel and semantic sub-tracks of AIGI-Now.

6 Ablation Study

We ablate five variants under an identical protocol (frozen MetaCLIP-2, layer 32, k=100, GenImage SD v1.4 train reference): forward $[\mathrm{CLS}]$ token, DPG-PatchQuad, and three rotation-based gradient aggregations—DPG-AbsMean, DPG-Anchor, and DPG-Probe (Ours).

Table 3: Ablation Performance on In-the-Wild Benchmarks.

Method	Chameleon			WildRF			CommunityAI			WebAIG-25			GPT2	Avg.
	Real	Fake	Avg.	Real	Fake	Avg.	Real	Fake	Avg.	Real	Fake	Avg.	Fake	
[CLS] token	0.973	0.468	0.720	0.371	0.888	0.629	0.959	0.658	0.808	0.799	0.709	0.754	0.393	0.661
DPG-PatchQuad	0.688	0.766	0.727	0.513	0.955	0.734	0.649	0.800	0.725	0.724	0.816	0.770	0.614	0.714
DPG-AbsMean	**0.979**	0.860	0.920	**0.597**	0.973	**0.785**	**0.979**	0.915	0.947	**0.808**	0.919	0.863	0.595	0.822
DPG-Anchor	0.966	0.922	0.944	0.485	0.983	0.734	0.968	0.971	0.970	0.775	0.941	0.858	0.696	0.840
DPG-Probe (Ours)	0.962	**0.937**	**0.950**	0.475	**0.983**	0.729	0.965	**0.978**	**0.971**	0.799	**0.942**	**0.870**	**0.704**	**0.845**

6.1 In-the-Wild Benchmarks

Table 3 shows that the forward $[\mathrm{CLS}]$ baseline achieves high real accuracy but weak fake detection, confirming the semantic trap discussed in Sec. 2. DPG-PatchQuad remains clearly below the rotation family (Avg. 0.714), validating that T must preserve identity. Among rotation-based variants, DPG-Probe attains the best overall average (0.845) with the strongest fake-side scores on the hardest targets (CommunityAI Fake 0.978, WebAIG-25 Fake 0.942, GPT2 0.704).

6.2 Out-of-Semantic-Domain Benchmarks

Table 4 reveals the same trends under semantic shift. The $[\mathrm{CLS}]$ token collapses on AnimeBench (0.306), while DPG-PatchQuad underperforms on dual-class suites such as GenImage (Avg. 0.768)

Table 4: Ablation Performance on Out-of-Semantic-Domain Benchmarks.

Method	GenImage++	AnimeBench	CoreBench	GenImage			AIGIHolmes			AIGI-Now Pixel			AIGI-Now Semantic			Avg.
				Real	Fake	Avg.	Real	Fake	Avg.	Real	Fake	Avg.	Real	Fake	Avg.	
[CLS] token	0.930	0.306	0.931	0.974	0.700	0.837	0.971	0.693	0.832	0.971	0.917	0.944	0.990	0.391	0.690	0.782
DPG-PatchQuad	0.897	0.856	0.918	0.804	0.731	0.768	0.839	0.793	0.816	0.838	0.907	0.873	0.961	**0.923**	**0.942**	0.867
DPG-AbsMean	0.956	0.875	0.919	0.978	0.708	0.843	0.968	0.894	0.931	0.959	0.911	0.935	**0.994**	0.762	0.878	0.905
DPG-Anchor	**0.966**	0.926	0.953	0.979	0.790	0.885	**0.976**	**0.939**	**0.957**	**0.974**	**0.941**	**0.958**	0.950	0.891	0.921	0.938
DPG-Probe (Ours)	0.964	**0.950**	**0.953**	**0.979**	**0.792**	**0.886**	0.975	0.937	0.956	0.972	0.940	0.956	0.944	0.912	0.928	**0.942**

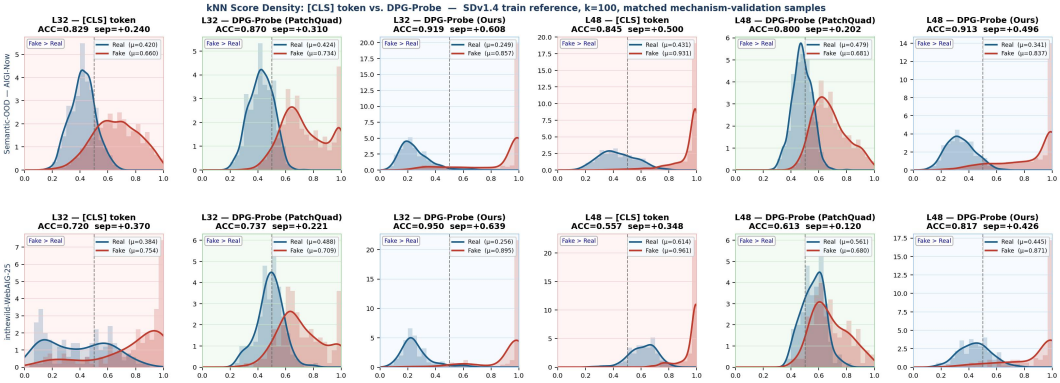


Figure 6: Gradient geometry inconsistency: Real vs. Fake (no labels). Distributions at Layer 32 (left) and Layer 48 (right) for a semantic-OOD target (AIGI-Now, top) and an in-the-wild target (WebAIG-25, bottom).

and AIGIHolmes (Avg. 0.816). DPG-Anchor is competitive on several benchmarks (AIGIHolmes Avg. 0.957), but DPG-Probe achieves the highest overall average (0.942) and wins the most challenging cases, including AnimeBench (0.950), CoreBench (0.953), and AIGI-Now Semantic (Avg. 0.928), demonstrating that dual-preserving rotation together with differential gradient purification is essential for cross-domain provenance retrieval.

6.3 Ablation of Gradient Direction Inconsistency

To directly verify why our multi-view rotation strategy outperforms conventional consistency paradigms, we record and compare their gradient direction inconsistency scores.

As qualitatively demonstrated in Fig. 6, conventional contrastive strategies like PatchQuad can partially mitigate semantic overfitting and exhibit larger gradient geometry inconsistencies at Layer 32 compared to Layer 48. Different spatial views exhibit drastic semantic variations. They remain heavily susceptible to an implicit supervision paradigm. In contrast, our dual-preserving gradient preserves the holistic semantic identity while probing pure geometric consistency. This allows rotation-induced gradient geometry inconsistency to achieve a much cleaner separation between real and fake distributions at intermediate layers (e.g., Layer 32), significantly outperforming the patch-based alternative.

7 Conclusion

We reframed AIGI detection as a provenance problem rather than semantic classification, and proposed DPG-Probe, a training-free method that extracts dual-preserving CLS gradients via a single backward pass and performs nearest-neighbor retrieval over a compact reference bank. Across out-of-semantic-domain and in-the-wild benchmarks, DPG-Probe consistently outperforms prior detectors in cross-generator generalization and real-world robustness. We hope this view of provenance as invariance failure under dual-preserving transformation provides a stable basis for future training-free AIGI forensics.

AI USE STATEMENT

In this work, we used generative AI tools for refining and polishing the text to improve clarity, grammar, and readability. We have not used generative AI tools for writing core technical sections, generating research ideas, writing code, or performing experimental evaluations, and data analysis and visual design are not applicable to this work. We have reviewed all AI-assisted work. Specifically, all AI-refined text was thoroughly checked and edited by the authors to ensure accuracy, proper terminology, and original intent without modifying any factual claims or technical descriptions. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

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

Ruoxin Chen, Jiahui Gao, Kaiqing Lin, Keyue Zhang, Yandan Zhao, Isabel Guan, Taiping Yao, and Shouhong Ding. Task-model alignment: A simple path to generalizable ai-generated image detection. *arXiv preprint arXiv:2512.06746*, 2025a.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, et al. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025b.

Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, Jim Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *Advances in Neural Information Processing Systems*, 38:48009–48036, 2026.

Feng Ding, Wenhui Yi, Yunpeng Zhou, Xinan He, Hong Rao, and Shu Hu. Fairness-aware deepfake detection: Leveraging dual-mechanism optimization. *arXiv preprint arXiv:2511.10150*, 2025.

Xinan He, Yue Zhou, Bing Fan, Bin Li, Guopu Zhu, and Feng Ding. Vlforgery face triad: Detection, localization and attribution via multimodal large language models. *Advances in Neural Information Processing Systems*, 38:163010–163044, 2026.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2405–2414, 2025a.

Ouxiang Li, Yuan Wang, Xinting Hu, Huijuan Huang, Rui Chen, Jiarong Ou, Xin Tao, Pengfei Wan, Xiaojuan Qi, and Fuli Feng. Easier painting than thinking: Can text-to-image models set the stage, but not direct the play? *arXiv preprint arXiv:2509.03516*, 2025b.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? *Advances in Neural Information Processing Systems*, 38, 2026.

Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16815–16825, 2024.

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.

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.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5052–5060, 2024a.

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

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.

Jiazhen Yan, Ziqiang Li, Fan Wang, Boyu Wang, Ziwen He, and Zhangjie Fu. Dgs-net: Distillation-guided gradient surgery for clip fine-tuning in ai-generated image detection. *arXiv preprint arXiv:2511.13108*, 2025a.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. In *International Conference on Learning Representations*, volume 2025, pp. 70702–70720, 2025b.

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

Kidus Zewde, Simiao Ren, Xingyu Shen, Jiaqi Wu, Yuchen Zhou, Tommy Duong, Zikang Zhang, Ethan Traister, and Kewen Xie. Gpt-image-2 in the wild: A twitter dataset of self-reported ai-generated images from the first week of deployment. *arXiv preprint arXiv:2604.25370*, 2026.

Yue Zhou, Xinan He, KaiQing Lin, Bing Fan, Feng Ding, and Bin Li. Breaking latent prior bias in detectors for generalizable aigc image detection. *Advances in Neural Information Processing Systems*, 38:30649–30679, 2026a.

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

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

Chenyang Zhu, Xing Zhang, Yuyang Sun, Ching-Chun Chang, and Isao Echizen. Animedl-2m: Million-scale ai-generated anime image detection and localization in diffusion era. In *Proceedings of the 1st on Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media*, pp. 45–54, 2025.

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

APPENDIX

STRUCTURE OF CONTENTS

- **Section A: Inference Overhead Analysis:**
  - Section A.1:Experimental Setup;
  - Section A.2: Comparison with Baseline Detectors;
  - Section A.3: Where Does the Overhead Come From?;
  - Section A.4: Accuracy–Efficiency Trade-off;
  - Section A.5: Deployment Cost at Scale;
- **Section B: WebAIG-25: A Leakage-Resistant In-the-Wild Benchmark**
- **Section C: Detailed Dataset Breakdown for Main Experiments:**
  - Section C.1: Fine-grained Evaluation on GenImage++;
  - Section C.2: Stylized Evaluation on Anime Benchmark;
  - Section C.3: Generalization Benchmarking on CoreBench;
  - Section C.4: Cross-domain Evaluation on GenImage;
  - Section C.5: Generalization on AIGIHolmes Suite;
  - Section C.6: Advanced Security Evaluation on AIGI-Now;
- **Section D: kNN Hyper-parameter Ablation:**
  - Section D.1: Experimental Setup;
  - Section D.2: Ablation on Neighborhood Size k;
  - Section D.3: Summary and Implications;
- **Section E: Robustness to Image Corruptions:**
  - Section E.1: Experimental Setup;
  - Section E.2: Results;
  - Section E.3: Discussion.

A INFERENCE OVERHEAD ANALYSIS

While DPG-Probe achieves strong cross-dataset generalization in the main experiments, its inference pipeline differs fundamentally from conventional supervised detectors. Instead of a single feed-forward pass through a frozen backbone followed by a lightweight linear head, DPG-Probe extracts *dual-preserving CLS-gradient features* via a multi-view forward–backward pass on MetaCLIP-2, and then scores each query image against a fixed reference bank using kNN. This design trades inference speed for label-free, reference-based adaptability. To make this trade-off explicit, we benchmark DPG-Probe against representative baselines from the Forensics-Comparison benchmark suite, covering three families: **(A)** lightweight CNN detectors, **(B)** CLIP-based linear classifiers, and **(C/D)** DPG-Probe variants with kNN or a linear probe on the same gradient features. All measurements are performed on a single NVIDIA L40 GPU using a unified compute-only protocol with synthetic 224×224 inputs; latency is estimated by repeated sampling rather than a full-dataset pass (Table 5).

A.1 EXPERIMENTAL SETUP

Protocol. We profile each method under identical hardware and input conditions. Baselines run with batch size $B{=}32$ and require only a single forward pass. DPG-Probe applies four rotation views ($0°, 90°, 180°, 270°$), computes a self-consistency loss, and performs one backward pass to obtain CLS-gradient features at layer 32 with `diff` aggregation. At inference time, DPG-Probe uses kNN with $k{=}100$ over the *full* GenImage SD v1.4 train reference pool ($\sim$162,000 real + $\sim$162,000 fake). For the kNN *head* micro-benchmark only, we subsample to 500 images per class ($1{,}000$ total) so that head latency can be isolated from gradient extraction; this subsample does *not* affect the reported gradient-extraction or end-to-end timings. Each configuration is warmed up for 20 iterations and timed over 50 iterations; we report mean per-image latency, 95th-percentile latency, throughput, and peak GPU memory. Accuracy values are taken from existing full-dataset evaluations (main paper: layer 32, $k{=}100$) and are not re-computed during profiling.

DPG-Probe configuration. Table 6 summarizes the inference hyper-parameters, which match the best-performing setup in the main paper (layer 32, $k{=}100$, `diff` aggregation).

Table 5: **Inference overhead measurement protocol.** All methods share the same GPU, resolution, and compute-only input mode. Latency is a sampling-based estimate (20 warmup + 50 timed iterations).

Item	Setting
GPU	NVIDIA L40 (46 GB, CUDA device 0)
Input resolution	224×224
Measurement mode	Compute-only (synthetic tensors, no I/O)
Warmup / timed iterations	20 / 50
Baseline batch size	32
DPG-Probe batch size	1 and 4
DPG-Probe backbone	MetaCLIP-2-worldwide-giant
DPG-Probe geometry	4-way rotation
DPG-Probe gradient aggregation	`diff`
DPG-Probe scoring	Layer 32, kNN (k=100) or linear probe
Reference bank (production)	SD v1.4 train, full pool ($\sim$324,000 images)
Reference bank (kNN head timing)	500 per class subsample (1,000 total)
Accuracy	Mean acc. from existing full-dataset eval (read-only)

Table 6: **DPG-Probe inference configuration.** Matches the best setup from the main experiments.

Hyper-parameter	Value
Vision encoder	`facebook/metaclip-2-worldwide-giant`
Feature type	CLS-token gradient (`cls_grad`)
Geometry augmentation	4-way rotation
Gradient aggregation	`diff`
Scoring layer	32
Scoring head	kNN (k=100) or linear probe
Reference bank	SD v1.4 train, full pool ($\sim$324,000 images)
Feature dimension	1,664

A.2 Comparison with Baseline Detectors

Table 7 reports the full comparison. **Group A** (ResNet50, NPR, FreqNet, SAFE) represents efficient CNN-based detectors whose per-image latency stays below 5 ms and peak VRAM below 3.2 GB. Among them, ResNet50 and NPR are the fastest ($\sim$0.28 ms/image, $\sim$3,500 img/s), reflecting the low cost of a single ResNet-50 forward pass with a one-layer head. **Group B** (CLIPDetection, Effort, DGS-Net) uses CLIP-L/14 backbones with linear or sigmoid heads; all three run at $\sim$8 ms/image ($\sim$125 img/s) with 1.7–2.9 GB peak VRAM—roughly $30\times$ slower than Group A but still well within real-time batch processing on a single GPU.

In contrast, **DPG-Probe (kNN)** at B=1 requires **357.7 ms/image** (2.8 img/s) with 17.5 GB peak VRAM. This is approximately **45**$\times$ slower than the strongest CLIP-based baseline (Effort, 7.9 ms) and over **1,200**$\times$ slower than lightweight CNN detectors. Increasing the batch size to B=4 reduces per-image latency to 320.8 ms (-10.3%), at the cost of higher peak memory (28.0 GB). Replacing kNN with a linear probe (Group D) does *not* materially change end-to-end latency (361.5 ms at B=1), confirming that the bottleneck lies in gradient feature extraction, not in the scoring head.

Crucially, DPG-Probe delivers a **mean accuracy of 83.7%** on the benchmark evaluation used in the main paper (layer 32, k=100), outperforming Effort (69.5%) by **+14.2 percentage points** and NPR (52.1%) by **+31.6 pp**. (Baseline Acc. columns in Table 7 come from the Forensics-Comparison suite.) The overhead is therefore not an implementation inefficiency but an inherent consequence of the dual-preserving gradient probe and the larger MetaCLIP-2 backbone.

A.3 Where Does the Overhead Come From?

To identify the dominant cost, we decompose DPG-Probe end-to-end latency into *gradient extraction* (4-view forward + backward on MetaCLIP-2) and *scoring* (kNN or linear probe). As shown in Table 8, gradient extraction accounts for **98.6%** of total latency at B=1 (352.7 ms out of 357.7 ms),

Table 7: **Inference overhead comparison between DPG-Probe and representative baselines.** Group A: lightweight CNN; Group B: CLIP + linear/sigmoid; Group C: DPG-Probe + kNN; Group D: DPG-Probe + linear probe. Acc. is mean accuracy (%) over all benchmark datasets (read-only eval).

Grp.	Method	Backbone / Head	Latency (ms/img)	p95 (ms)	Throughput (img/s)	VRAM (MB)	Acc. (%)
A	ResNet50	ResNet50 + sigmoid	0.28	0.28	3544.49	369	–
	NPR	ResNet50 + sigmoid	0.28	0.28	3542.40	369	52.1
	FreqNet	FreqNet + sigmoid	4.75	4.76	210.52	3170	56.7
	SAFE	ResNet50 + softmax	0.31	0.31	3199.21	351	53.9
B	CLIPDetection	CLIP + sigmoid	8.22	8.31	121.59	2880	62.6
	Effort	CLIP-L/14 + linear	7.93	8.07	126.10	1699	69.5
	DGS-Net	CLIP-L/14 + linear	8.09	8.21	123.56	2883	44.4
C	**DPG-Probe (kNN)**	MetaCLIP-2-giant + kNN	**357.74**	**362.14**	**2.80**	**17456**	**83.7**
	DPG-Probe (kNN, B=4)	MetaCLIP-2-giant + kNN	320.80	323.16	3.12	28013	83.7
D	DPG-Probe (Linear)	MetaCLIP-2-giant + linear	361.53	366.31	2.77	17456	–
	DPG-Probe (Linear, B=4)	MetaCLIP-2-giant + linear	319.15	321.81	3.13	28013	–

while the kNN head contributes only **1.4 ms** (0.38%) under the 1,000-image profiling subsample (k=100). With the full ∼324,000-image production reference, kNN scoring would add non-negligible cost; we therefore treat gradient extraction as the dominant bottleneck and leave full-scale kNN profiling to future work. Switching to a linear probe reduces the head cost to 0.08 ms (0.02%) without meaningfully changing the total. Because the two stages are profiled independently, their sum can be slightly lower than the end-to-end timing; the E2E column should be treated as the authoritative per-query cost. This confirms that further speed-ups must target the gradient extraction stage (e.g., fewer views, a smaller backbone, mixed-precision, or layer-selective backprop), not the kNN classifier.

Table 9 quantifies the relative cost. Against the most accurate CLIP-based baseline (Effort, 69.5% acc.), DPG-Probe is **45.1**× slower but gains **+14.2 pp** in accuracy. Against NPR, the slowdown exceeds **1,200**×, with a **+31.6 pp** accuracy gain. Even compared to FreqNet—the slowest Group A method—DPG-Probe is **75.3**× slower while improving accuracy by **+27.0 pp**.

Table 8: **Latency breakdown of DPG-Probe.** Gradient extraction (4-view forward + backward) dominates; the scoring head contributes <0.5%. Grad. extract and scoring head are profiled independently; E2E is measured jointly.

Variant	Grad Extract (ms/img)	Scoring Head (ms/img)	E2E Total (ms/img)	Head Share (%)
DPG-Probe (kNN, B=1)	352.75	1.36	357.74	0.38
DPG-Probe (Linear, B=1)	359.87	0.08	361.53	0.02
DPG-Probe (kNN, B=4)	315.01	1.38	320.80	0.43
DPG-Probe (Linear, B=4)	315.76	0.08	319.15	0.02

Table 9: **Relative inference cost of DPG-Probe (kNN, B=1) w.r.t. baselines.** Slowdown = DPG-Probe latency / baseline latency. Acc. gain = DPG-Probe acc. − baseline acc. (pp).

Baseline	Latency (ms/img)	Slowdown (×)	Acc. (%)	Acc. Gain (pp)
ResNet50	0.28	1268.5×	–	–
NPR	0.28	1267.2×	52.1	+31.6
SAFE	0.31	1144.4×	53.9	+29.8
FreqNet	4.75	75.3×	56.7	+27.0
DGS-Net	8.09	44.2×	44.4	+39.3
CLIPDetection	8.22	43.5×	62.6	+21.1
Effort	7.93	45.1×	69.5	+14.2
DPG-Probe (kNN)	**357.74**	**1.0**×	**83.7**	–

A.4 Accuracy–Efficiency Trade-off

Table 10 summarizes the accuracy–efficiency trade-off from a deployment perspective. Conventional detectors cluster in the low-latency, moderate-accuracy regime: NPR and SAFE achieve $\sim$0.3 ms latency but only 52–54% mean accuracy, while Effort reaches 69.5% at $\sim$8 ms. DPG-Probe occupies a distinct point in this space: **83.7% accuracy at 357.7 ms/image**. Although its Acc./Latency ratio is low (0.23 vs. 8.76 for Effort), this is an expected consequence of prioritizing *cross-dataset generalization via a fixed reference bank* over raw inference speed. In settings where detection accuracy across diverse generators matters more than real-time throughput—e.g., offline forensic screening or batch auditing—the additional cost is justified by the substantial accuracy gain.

Table 10: **Accuracy–latency trade-off.** DPG-Probe trades speed for cross-dataset mean accuracy.

Method	Mean Acc. (%)	Latency (ms/img)	Acc./Latency
DGS-Net	44.4	8.09	5.49
NPR	52.1	0.28	185.36
SAFE	53.9	0.31	173.87
FreqNet	56.7	4.75	11.94
CLIPDetection	62.6	8.22	7.62
Effort	69.5	7.93	8.76
DPG-Probe (kNN)	**83.7**	**357.74**	**0.23**

A.5 Deployment Cost at Scale

One-time reference construction. DPG-Probe requires a reference bank to be built offline before online inference. As reported in Table 11, extracting gradient features for the full SD v1.4 train pool ($\sim$324,000 images) would take an estimated $\sim$**34 hours** on the same L40 GPU ($\sim$375 ms/image $\times$ $\sim$324,000 images). The table also reports a 1,000-image subsample timing used to sanity-check the per-image extraction rate. This one-time cost is amortized over all subsequent queries and is *not* included in the per-query latencies above.

Large-scale inference. Table 12 extrapolates per-image latency to a 100,000-image workload. While all baselines complete within minutes, DPG-Probe (kNN, $B{=}1$) requires an estimated **9.9 hours**. Using $B{=}4$ reduces this to 8.9 hours (-11%). For practical deployment, we therefore recommend: (i) pre-building and caching the reference bank; (ii) using $B{=}4$ or higher where GPU memory permits; and (iii) reserving DPG-Probe for accuracy-critical offline pipelines rather than latency-sensitive online filtering.

Table 11: **One-time offline cost of building the DPG-Probe reference bank.** Amortized over all online queries; not included in per-query latency above.

Item	Value
Reference dataset	GenImage / stable_diffusion_v1.4 / train
Reference size (production)	$\sim$324,000 images ($\sim$162,000 per class)
Per-image extraction time (estimated)	375.64 ms
Estimated total build time (production)	$\sim$34 h
1,000-image subsample build time	6.3 min (profiling sanity check)
kNN index size (production)	$\sim$324,000 $\times$ 1,664($L2-normalized$)

A.6 Summary

This overhead analysis highlights three key findings:

1. **Gradient extraction dominates.** Over 98% of DPG-Probe latency comes from the 4-view forward–backward pass on MetaCLIP-2; the kNN scoring head adds less than 2 ms per image.

Table 12: **Extrapolated inference cost for** 100,000 **images.** Baselines use B=32; DPG-Probe uses B=1 unless noted.

Method	Latency (ms/img)	Est. time for 100K images
ResNet50	0.28	0.008 h (≈ 28 s)
NPR	0.28	0.008 h (≈ 28 s)
SAFE	0.31	0.009 h (≈ 31 s)
FreqNet	4.75	0.132 h (≈ 8 min)
Effort	7.93	0.220 h (≈ 13 min)
CLIPDetection	8.22	0.228 h (≈ 14 min)
DGS-Net	8.09	0.225 h (≈ 14 min)
DPG-Probe (kNN)	**357.74**	**9.94 h**
DPG-Probe (kNN, B=4)	320.80	8.91 h

2. **The cost buys generalization.** Compared to Effort—the strongest CLIP-based baseline—DPG-Probe is ∼45× slower but improves mean accuracy from 69.5% to 83.7% (+14.2 pp) without per-dataset retraining.
3. **Deployment is feasible for offline use.** With a one-time offline reference build (∼34 h for the full pool) and batch size B=4, DPG-Probe can process ∼3.1 images/second on a single L40 GPU, making it suitable for batch forensic screening where accuracy is paramount.

Future work may reduce overhead via mixed-precision gradient extraction, view pruning, or distillation into a single-pass student model, while preserving the reference-based generalization that distinguishes DPG-Probe from conventional linear detectors.

B WebAIG-25: A Leakage-Resistant In-the-Wild Benchmark

Existing in-the-wild benchmarks may still overlap with the web-scale pre-training corpora of modern vision foundation models (VFMs). To provide a stringent *containment test* against pre-training data leakage, we introduce **WebAIG-25**, a custom blind benchmark whose real and synthetic portions are verifiably unseen by existing VFMs. WebAIG-25 comprises 5,937 images organized into two complementary splits—privately captured real photographs and post-2025 Midjourney artworks—stored under a unified directory structure (`WebAIG-25/real` and `WebAIG-25/fake`). Table 13 summarizes the split-level statistics.

B.1 Dataset Construction

Real split (`picsum_images`). The real portion consists of 2,965 privately captured, high-resolution photographs of natural scenes (landscapes, cityscapes, outdoor environments). All images are stored as JPEG and span three quality tiers (`high`, `ultra`, and `medium`) with resolutions ranging from 800×600 to 1920×1080. The most frequent resolutions are 1920×1080 (795 images), 1440×900 (773), and 1600×900 (747). Images were captured and exported in September 2025 and are completely insulated from any public web-crawled pre-training corpora.

Synthetic split (`midjourney_images`). The synthetic portion contains 2,972 AI-generated artworks scraped from the `r/midjourney` subreddit, restricted to posts published strictly after the year 2025 (August 13–September 10, 2025; 447 unique posts). Each post is accompanied by a JSON metadata file recording the Reddit `post_id`, title, author, creation timestamp, permalink, upvote score, and source image URLs. We retain all gallery images from each qualifying post, yielding a median of 4 and a maximum of 20 images per post. Image formats include PNG (1,552), JPG (1,236), and JPEG (184), with resolutions ranging from 256×256 to 5824×4928 pixels (median width/height: 1024×1232). Because all content was posted after the training cut-off dates of mainstream VFMs, this split serves as a decisive test against synthetic-data leakage.

Visual samples. Figure 7 (main paper, Fig. 5) shows representative images from both splits. The real split exhibits diverse natural-scene photography at multiple resolution tiers, while the synthetic split covers a wide range of Midjourney art styles including portraits, fantasy scenes, cyberpunk art,

Table 13: Summary statistics of the WebAIG-25 benchmark. Real and fake splits are also referenced as `picsum_images` and `midjourney_images` in our evaluation code.

Attribute	**Real (`picsum`)**	**Fake (`midjourney`)**
# Images	2,965	2,972
# Source units	–	447 Reddit posts
File formats	JPEG	PNG / JPG / JPEG
Resolution range	800^2–1920^2	256^2–5824^2
Median resolution	1600×900	1024×1232
Median file size	0.13 MB	0.44 MB
Collection period (2025)	Sep. 9–10	Aug. 13–Sep. 10
Disk usage	0.43 GB	2.17 GB
Total	5,937 images (∼2.6 GB)	

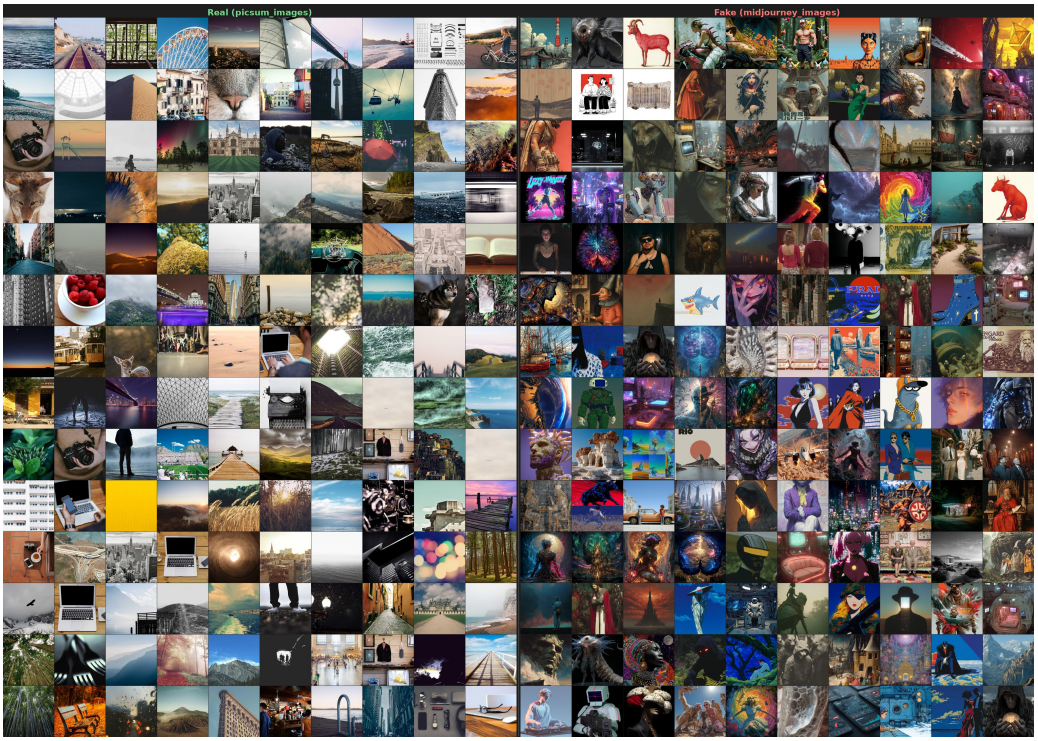


Figure 7: **Representative samples from WebAIG-25. Left:** `picsum_images` (real, `real/`). **Right:** `midjourney_images` (fake, `fake/`). Randomly sampled visualization subset; center-cropped to square tiles.

and product visualizations. Notably, the `flux_amateur` panel in Fig. 5 belongs to our GenImage++ benchmark (Sec. C.1) and is included for visual comparison only; it is *not* part of WebAIG-25.

B.2 Evaluation Protocol

In our codebase, WebAIG-25 supports three evaluation modes: (i) `WebAIG-25` as a dual-class benchmark with `real/` and `fake/` subdirectories; (ii) `picsum_images` as a real-only subset (label $= 0$); and (iii) `midjourney_images` as a fake-only subset (label $= 1$). For kNN score-density and mechanism-validation analyses, the two single-class subsets are merged into a paired benchmark denoted `picsum+midjourney`. Unless otherwise noted, dual-class accuracy is reported as $\mathrm{Acc} = (\mathrm{acc}_{\mathrm{real}} + \mathrm{acc}_{\mathrm{fake}})/2$.

B.3 DISCUSSION

WebAIG-25 complements existing in-the-wild suites by enforcing temporal and provenance constraints on *both* classes: the real images are privately captured (not web-crawled), and the synthetic images are community-generated content posted after VFM training cut-offs. This dual containment makes WebAIG-25 a particularly stringent benchmark for evaluating whether a detector relies on memorized pre-training artifacts rather than genuine forensic signals. In our main in-the-wild evaluation (Table 1), WebAIG-25 achieves $\mathrm{Acc}_{\mathrm{real}} = 0.799$ and $\mathrm{Acc}_{\mathrm{fake}} = 0.942$ ($\mathrm{Avg.} = 0.870$), outperforming all baselines on this benchmark.

C DETAILED DATASET BREAKDOWN FOR MAIN EXPERIMENTS

In this section, we present the comprehensive, fine-grained performance breakdown of all compared forensic detectors across individual generator subsets. These detailed results provide a deeper multi-view complement to the macro-averaged results reported in Table 1 and Table 2 of the main text, further validating the robustness and cross-domain generalizability of our training-free `DPG-Probe` framework.

C.1 FINE-GRAINED EVALUATION ON GENIMAGE++

Table 14 provides the subset-level Accuracy (%) on our proposed GenImage++ benchmark. This benchmark includes challenging iterations of advanced synthesis engines such as Flux variants and Stable Diffusion models. Notably, on the highly deceptive *Flux Amateur* subset, standard semantic-dominant classifiers (e.g., CNNSpot, GramNet) exhibit a near-complete performance collapse ($\leq 6.1\%$). In contrast, Ours consistently demonstrates strong resilience across all distributions, reaching 82.4% on *Flux Amateur* and securing the best average accuracy of 96.1%, with leading performance on Flux (99.6%), Flux Multi (99.6%), Flux Photo (100.0%), Flux Real (100.0%), and SD3 Real (99.5%).

Table 14: Performance (Accuracy %) of detectors on our proposed GenImage++ benchmark subsets. The best results within each subset are highlighted in bold font.

Model	Flux	Flux Amateur	Flux Multi	Flux Photo	Flux Real	SD1.5 Multi	SDXL Multi	SD3	SD3 Photo	SD3 Real	Avg
Semantic-Dominant Projections											
CNNSpot	0.461	0.021	0.079	0.027	0.117	0.940	0.199	0.516	0.026	0.246	0.221
FreqNet	0.949	0.811	0.618	0.958	0.800	0.968	0.794	0.921	**0.965**	0.760	0.874
GramNet	0.652	0.061	0.151	0.179	0.225	0.905	0.400	0.566	0.159	0.274	0.327
UnivFD	0.929	0.159	0.856	0.864	0.917	0.889	0.857	0.917	0.744	0.896	0.820
LaDeDa	0.365	0.277	0.372	0.296	0.559	0.784	0.387	0.266	0.612	0.496	0.464
Effort	0.971	0.757	0.966	0.926	0.991	0.992	0.979	**0.990**	**0.965**	0.993	0.957
AIDE	0.940	0.743	0.950	0.946	0.947	0.975	0.971	0.910	0.887	0.924	0.930
DGS-Net	0.446	0.026	0.353	0.175	0.400	0.618	0.615	0.490	0.170	0.524	0.350
Micro-Artifact Anomalies											
NPR	0.981	0.649	0.862	0.978	0.969	0.974	0.645	0.897	0.921	0.923	0.890
SAFE	0.913	**0.974**	0.838	0.900	0.969	0.805	0.760	0.841	0.932	0.880	0.872
DDA	0.912	0.697	0.891	0.936	0.966	0.982	**0.997**	0.976	**0.965**	0.992	0.944
OMAT	0.965	0.281	0.925	0.976	0.976	**1.000**	0.991	0.982	0.903	0.988	0.899
Ours	**0.996**	0.824	**0.996**	**1.000**	**1.000**	0.922	0.995	0.973	0.906	**0.995**	**0.961**

C.2 STYLIZED EVALUATION ON ANIME BENCHMARK

Table 15 investigates the fake detection accuracy on non-photorealistic artistic domains across 19 distinct unseen anime generator configurations. Due to the massive domain shift from natural web-scraped images to stylized illustrations, baseline networks parameter-tuned on photorealistic datasets completely break down, often yielding near-blind detection rates ($\leq 10\%$). Conversely, by purifying structural forgery tokens orthogonal to contextual textures, Ours demonstrates an unprecedented zero-shot capability, securing the best average accuracy of 95.0% and achieving the best or tied-best performance on all 19 unseen generators (e.g., 96.4% on SD 1.5 and 97.5% on Pony).

Table 15: Performance on Anime Benchmark. Fake accuracy values are detailed for 19 unseen generative tools. Best results are marked in bold.

Method	SD 1.4	SD 1.5	SD 1.5 Hyper	SD 1.5 LCM	SD 2.0	SD 2.0 768	SD 2.1	SD 2.1 768	SDXL 0.9	SDXL 1.0	SDXL 1.0 LCM	SDXL Hyper	SDXL Light	SDXL Turbo	Pony	Illus	Flux.1 D	Flux.1 S	Other	Avg.
Semantic-Dominant Projections																				
CNNSpot	0.064	0.054	0.071	0.333	0.136	0.000	0.000	0.074	0.182	0.117	0.125	0.091	0.179	0.258	0.032	0.052	0.059	0.073	0.055	0.057
FreqNet	0.064	0.091	0.000	0.603	0.295	0.000	0.000	0.059	0.000	0.075	0.750	0.091	0.400	0.178	0.048	0.201	0.067	0.018	0.049	0.086
Gram-Net	0.028	0.019	0.000	0.295	0.091	0.000	0.000	0.044	0.000	0.015	0.062	0.030	0.046	0.044	0.008	0.020	0.008	0.018	0.012	0.017
UnivFD	0.872	0.819	0.714	0.756	0.773	**1.000**	0.818	0.441	0.727	0.632	0.750	0.848	0.607	0.467	0.777	0.739	0.382	0.436	0.820	0.782
LaDeDa	0.009	0.028	0.000	0.218	0.068	0.000	0.000	0.000	0.000	0.009	0.000	0.030	0.079	0.022	0.005	0.011	0.004	0.000	0.012	0.020
Effort	0.761	0.776	0.571	0.833	**0.955**	**1.000**	0.636	0.618	0.682	0.722	0.875	0.879	0.882	0.831	0.650	0.634	0.627	0.582	0.761	0.738
AIDE	0.018	0.073	0.071	0.474	0.250	0.000	0.000	0.015	0.000	0.043	0.062	0.091	0.146	0.124	0.044	0.117	0.031	0.018	0.043	0.066
DGS-Net	0.018	0.025	0.000	0.269	0.068	0.500	0.000	0.279	0.000	0.022	0.188	0.061	0.236	0.067	0.009	0.014	0.024	0.018	0.013	0.022
Micro-Artifact Anomalies																				
NPR	0.028	0.072	0.000	0.615	0.273	0.000	0.000	0.000	0.000	0.025	0.125	0.091	0.314	0.151	0.031	0.092	0.011	0.000	0.029	0.060
SAFE	0.037	0.080	0.000	0.577	0.318	0.000	0.000	0.059	0.000	0.034	0.250	0.121	0.339	0.142	0.034	0.108	0.011	0.000	0.035	0.067
DDA	0.688	0.678	0.286	0.782	0.841	0.500	0.773	0.618	0.455	0.606	0.500	0.303	0.586	0.582	0.628	0.616	0.264	0.327	0.684	0.651
OMAT	0.202	0.246	0.071	0.859	0.614	**1.000**	0.318	0.441	0.477	0.267	**1.000**	0.242	0.471	0.467	0.170	0.357	0.178	0.145	0.258	0.240
Ours	**0.954**	**0.964**	**0.929**	**1.000**	**0.955**	**1.000**	**1.000**	**0.868**	**0.955**	**0.943**	**1.000**	**1.000**	**0.957**	**0.938**	**0.975**	**0.939**	**0.910**	**0.800**	**0.957**	**0.950**

C.3 Generalization Benchmarking on CoreBench

Table 16 highlights the individual breakdown across modern next-generation architectures grouped under the CoreBench series. This evaluation suite focuses heavily on bleeding-edge autoregressive frameworks and multi-modal transformers released after 2024. While heavily parameterized discriminators like DDA or AIDE manage to hit peak metrics on specific architectures (e.g., Gemini, BLIP models), they reveal an extreme inconsistency when encountering architectural variance (e.g., DDA dropping to 24.8% on Flux-Krea). Ours strikes the most reliable performance balance, leading the absolute average across all 20 advanced generator splits (95.3%), with consistently strong results on FLUX variants (e.g., 95.8% on FLUX.1-dev and 92.4% on FLUX.1-Krea-dev).

Table 16: Performance on CoreBench. Fake accuracy metrics across individual modern open-source and closed-source generative pipelines. Best results are marked in bold.

Method	FLUX.1 dev	FLUX.1 Krea-dev	FLUX.1 schnell	FLUX.2 dev	FLUX.2 klein-4B	FLUX.2 klein-9B	BAGEL	BLIP3o 4B	BLIP3o 8B	Gemini Flash2.0	GPT-4o	GPT-Image 1.5	GoT-R1 7B	HiDream I1	Hunyuan Image3.0	Infinity 8B	Janus Pro-1B	Janus Pro-7B	Nano Banana-2	OmniGen2 7B	Avg.
Semantic-Dominant Projections																					
CNNSpot	0.120	0.086	0.187	0.248	0.053	0.085	0.147	0.055	0.056	0.178	0.026	0.087	0.352	0.216	0.099	0.230	0.263	0.274	0.055	0.310	0.156
FreqNet	0.863	0.754	0.830	0.873	0.823	0.838	0.881	0.672	0.560	0.809	0.907	0.813	0.954	0.759	0.910	0.887	0.897	0.968	0.779	0.732	0.826
Gram-Net	0.220	0.151	0.274	0.299	0.134	0.147	0.279	0.052	0.038	0.390	0.158	0.139	0.450	0.195	0.074	0.252	0.612	0.692	0.118	0.380	0.253
UnivFD	0.779	0.722	0.746	0.768	0.688	0.699	0.816	0.829	0.816	0.757	0.765	0.733	0.912	0.589	0.793	0.889	0.778	0.724	0.704	0.775	0.764
LaDeDa	0.352	0.137	0.356	0.397	0.251	0.270	0.297	0.203	0.229	0.361	0.381	0.561	0.244	0.182	0.768	0.442	0.397	0.471	0.499	0.299	0.355
Effort	0.885	**0.951**	0.931	0.935	0.944	0.930	0.941	0.893	0.880	0.935	0.938	**0.965**	0.994	0.876	0.980	0.974	0.995	0.995	0.969	0.850	0.938
AIDE	0.902	0.887	0.791	0.921	0.864	0.906	0.872	0.970	0.953	0.819	**0.984**	0.903	0.259	0.391	0.989	0.975	0.194	0.186	0.970	0.784	0.776
DGS-Net	0.308	0.294	0.318	0.344	0.131	0.153	0.415	0.378	0.358	0.401	0.393	0.291	0.835	0.135	0.338	0.479	0.949	0.911	0.198	0.375	0.400
Micro-Artifact Anomalies																					
NPR	0.929	0.640	0.894	0.863	0.785	0.701	0.920	0.374	0.388	0.863	0.748	0.702	0.875	0.840	0.659	0.877	0.908	0.956	0.730	0.900	0.778
SAFE	0.898	0.778	0.843	0.921	0.863	0.874	0.882	0.498	0.365	0.815	0.930	0.834	0.948	0.911	0.971	0.962	0.936	0.958	0.889	0.863	0.847
DDA	0.933	0.248	**0.963**	0.899	0.950	**0.965**	0.950	**0.996**	**0.993**	**0.962**	0.727	0.619	0.997	0.808	**0.998**	**0.999**	0.995	0.995	**0.977**	**0.972**	0.897
OMAT	0.923	0.759	0.931	0.909	0.833	0.709	0.913	0.921	0.917	0.886	0.301	0.581	**1.000**	0.876	0.648	0.963	**1.000**	**1.000**	0.762	0.959	0.840
Ours	**0.958**	0.924	0.942	**0.956**	**0.955**	0.952	**0.987**	0.987	0.981	0.808	0.921	0.942	0.999	**0.970**	0.982	0.981	0.979	0.988	0.941	0.916	**0.953**

C.4 Cross-domain Evaluation on GenImage

Table 17 details the cross-generator classification results under the standard GenImage training configuration (trained on Stable Diffusion v1.4 and tested across completely unseen distribution suites). While standard structural artifact anomalies struggle when applied to specific diffusion branches (e.g., NPR scoring 56.7% on VQDM), Ours demonstrates strong cross-domain stability, yielding an average performance of 88.6% and reaching 97.3% on Midjourney while maintaining balanced accuracy across ADM (74.6%), VQDM (74.7%), and GLIDE (91.2%).

C.5 Generalization on AIGIHolmes Suite

Table 18 breaks down the forensic capabilities on advanced autoregressive and next-generation diffusion-transformer architectures compiled under the AIGIHolmes suite. On models using entirely unseen scaling receipts (e.g., Janus networks), baseline detectors experience a severe drop in performance, approaching a state of complete blindness ($\approx 43\% \sim 50\%$). By utilizing training-free backpropagated consistency descriptors, Ours successfully recovers these advanced fakes—achieving 85.3% on Janus and 96.6% on Janus-Pro-1B—while maintaining a strong suite average of 95.6% and leading on FLUX (98.3%).

Table 17: Performance on GenImage Benchmark. All detectors are trained on Stable Diffusion v1.4 and cross-evaluated on unseen generator splits. Best results are marked in bold.

Method	ADM	BigGAN	Midjourney	VQDM	GLIDE	SD-v1.4	SD-v1.5	Wukong	Avg.
Semantic-Dominant Projections									
CNNSpot	0.507	0.530	0.610	0.501	0.522	**0.998**	**0.996**	0.985	0.706
FreqNet	0.608	0.939	0.849	0.626	0.858	0.950	0.949	0.937	0.840
Gram-Net	0.576	0.591	0.665	0.528	0.747	0.989	0.988	0.960	0.756
UnivFD	0.571	0.839	0.795	0.644	0.842	0.959	0.958	0.926	0.817
LaDeDa	0.512	0.583	0.595	0.513	0.585	0.876	0.871	0.820	0.670
Effort	0.704	0.749	0.756	0.824	0.807	0.874	0.873	0.867	0.807
AIDE	0.602	0.648	0.813	0.694	0.514	0.959	0.957	0.954	0.768
DGS-Net	0.723	**0.969**	0.588	0.909	0.792	0.814	0.809	0.804	0.801
Micro-Artifact Anomalies									
NPR	0.619	0.849	0.850	0.567	0.844	0.996	0.994	0.984	0.838
SAFE	0.600	0.880	0.829	0.708	0.836	0.964	0.967	0.903	0.836
DDA	**0.880**	0.741	0.960	0.719	0.862	0.986	0.985	**0.986**	0.890
OMAT	0.837	**0.973**	0.903	**0.954**	**0.974**	0.974	0.973	0.975	**0.946**
Ours	0.746	0.829	**0.973**	0.747	0.912	0.972	0.972	0.936	0.886

Table 18: Generalization performance on the AIGIHolmes suite across individual advanced autoregressive and diffusion-transformer architectures. Best results are marked in bold.

Detector	FLUX	Infinity	Janus	Janus-Pro-1B	Janus-Pro-7B	LlamaGen	PixArt-XL	SD3.5-L	Show-o	VAR	Avg
Semantic-Dominant Projections											
CNNSpot	0.626	0.589	0.508	0.531	0.522	0.604	0.668	0.568	0.608	0.504	0.573
FreqNet	0.894	0.924	0.443	0.439	0.433	0.875	0.885	0.894	0.910	0.688	0.738
Gram-Net	0.696	0.604	0.491	0.491	0.491	0.627	0.627	0.701	0.771	0.557	0.606
UnivFD	0.868	0.898	0.575	0.674	0.575	0.915	0.925	0.898	0.911	0.597	0.784
LaDeDa	0.682	0.664	0.498	0.497	0.497	0.810	0.641	0.695	0.755	0.724	0.646
Effort	0.794	0.804	0.483	0.650	0.569	0.798	0.804	0.765	0.793	0.764	0.722
AIDE	0.892	0.921	0.447	0.448	0.448	0.933	0.935	0.942	0.893	0.651	0.751
DGS-Net	0.627	0.539	0.490	0.504	0.493	0.752	0.736	0.869	0.808	0.868	0.668
Micro-Artifact Anomalies											
NPR	0.968	0.983	0.495	0.495	0.495	0.947	0.899	0.937	**0.988**	0.730	0.794
SAFE	0.932	0.965	0.482	0.482	0.483	0.913	0.896	0.940	0.962	0.911	0.796
DDA	0.972	**0.989**	**0.985**	**0.993**	**0.987**	**0.993**	**0.994**	**0.970**	0.948	0.804	**0.963**
OMAT	0.947	0.955	0.651	0.756	0.641	0.967	0.969	0.957	0.969	**0.959**	0.877
Ours	**0.983**	0.986	0.853	0.966	0.961	0.988	0.986	0.968	0.985	0.886	0.956

C.6 Advanced Security Evaluation on AIGI-Now

Table 19 evaluates performance across the dual tracks of the AIGI-Now dataset, distinguishing between standard physical rendering fakes (*pix*) and high-level semantic fakes (*sem*). Notably, advanced commercial simulators elegantly optimize away low-level pixel artifacts, causing traditional baseline models to drop to near-blind classification accuracy ($\approx 45\% \sim 55\%$) on the semantic (*sem*) track. In contrast, by mapping dynamic responses within our unsupervised backpropagated manifold, Ours remains desensitized to high-level semantic offsets, consistently preserving strong performance on the *sem* track (e.g., 96.7% on FLUX-dev and 95.6% on gpt4o) and achieving the best overall average accuracy of 94.2%.

Table 19: Fine-grained generalization on the dual-track AIGI-Now benchmark. Performance breaks down into physical rendering (*pix*) and semantic (*sem*) sub-tracks across 9 unobserved models. Best results are marked in bold.

Detector	FLUX-dev		FLUX-kera		FLUX-kontext		FLUX-pro		gpt4o		jimeng		keling		minimax		Nano		Avg.
	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	pix	sem	
Semantic-Dominant Projections																			
CNNSpot	0.919	0.500	0.550	0.500	0.843	0.502	0.535	0.500	**0.990**	0.501	0.523	0.500	0.973	0.501	0.603	0.504	0.985	0.499	0.635
FreqNet	0.875	0.468	0.697	0.443	0.769	0.479	0.492	0.501	0.923	0.510	0.459	0.530	0.922	0.543	0.824	0.511	0.907	0.487	0.622
Gram-Net	0.933	0.528	0.667	0.522	0.864	0.555	0.608	0.569	0.763	0.508	0.508	0.503	0.955	0.554	0.717	0.521	0.905	0.528	0.651
UnivFD	0.542	0.579	0.514	0.538	0.492	0.545	0.516	0.544	0.529	0.532	0.475	0.504	0.631	0.539	0.501	0.539	0.501	0.516	0.531
LaDeDa	0.586	0.498	0.497	0.497	0.560	0.502	0.496	0.495	0.745	0.499	0.495	0.497	0.661	0.501	0.509	0.502	0.766	0.505	0.546
Effort	0.789	0.679	0.796	0.669	0.728	0.610	0.688	0.690	0.753	0.580	0.522	0.555	0.782	0.677	0.772	0.687	0.796	0.657	0.690
AIDE	**0.991**	0.590	0.504	0.569	**0.979**	0.806	0.601	0.538	0.747	0.518	0.639	0.514	**0.982**	0.554	0.514	0.541	**0.989**	0.518	0.672
DGS-Net	0.638	0.513	0.526	0.510	0.633	0.507	0.491	0.520	0.626	0.513	0.493	0.525	0.740	0.518	0.529	0.507	0.650	0.538	0.555
Micro-Artifact Anomalies																			
NPR	0.944	0.500	0.508	0.500	0.785	0.502	0.502	0.502	0.966	0.500	0.497	0.501	0.957	0.500	0.548	0.500	0.930	0.500	0.619
SAFE	0.903	0.490	0.532	0.492	0.831	0.494	0.521	0.486	0.977	0.488	0.509	0.486	0.960	0.487	0.590	0.491	0.961	0.484	0.621
DDA	0.916	0.512	0.594	0.499	0.827	0.529	0.766	0.550	0.923	0.654	0.870	0.654	0.961	0.646	0.833	0.505	0.816	0.562	0.695
OMAT	0.911	0.475	0.649	0.469	0.847	0.507	0.591	0.515	0.744	0.452	0.491	0.465	0.936	0.526	0.699	0.467	0.891	0.468	0.615
Ours	0.981	**0.967**	**0.966**	**0.931**	0.837	**0.835**	**0.960**	**0.938**	0.981	**0.956**	**0.968**	**0.942**	0.978	**0.939**	**0.949**	**0.901**	0.984	**0.945**	**0.942**

D kNN Hyper-parameter Ablation

The main experiments fix the kNN neighborhood size to $k{=}100$ and construct the reference bank from the full GenImage SD v1.4 training split. We ask how sensitive retrieval accuracy is to the choice of k under the same evaluation protocol used in the main paper: mean accuracy over six semantic out-of-domain (OOD) or six in-the-wild benchmarks. All variants use MetaCLIP-2 CLS-gradient features; we compare DPG-Probe (ours), DPG-Anchor, and DPG-AbsMean against a frozen MetaCLIP-2 $[\mathrm{CLS}]$ baseline ($k{=}10$).

D.1 Experimental Setup

Evaluation protocol. For each layer index, we extract L2-normalized features, retrieve k nearest neighbors from the fixed reference bank, and classify via the fake-neighbor fraction with a 0.5 threshold. Curves report the dataset-mean accuracy.

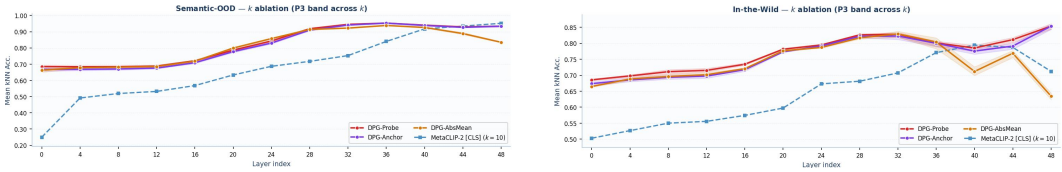
D.2 Ablation on Neighborhood Size k

Figure 8 overlays all DPG variants (plus the frozen $[\mathrm{CLS}]$ baseline at $k{=}10$) under the k-sweep band. Across both domain groups, three consistent patterns emerge.

Peak-layer accuracy is largely k-invariant on semantic-OOD. At L36, DPG-Probe varies by only **0.5 pp** (95.2% at $k{=}5 \rightarrow 95.7\%$ at $k{=}100$), while DPG-Anchor spans **0.6 pp** ($95.0\% \rightarrow 95.6\%$).

In-the-wild benchmarks are more k-sensitive, favoring larger k. At L32, DPG-Probe improves from **81.3%** ($k{=}5$) to **84.9%** ($k{=}100$), a **3.6 pp** swing; DPG-AbsMean shows an even wider **4.4 pp** gap. Smaller neighborhoods appear to overfit idiosyncratic reference neighbors on diverse wild domains, whereas $k{\geq}20$ stabilizes performance. Our default $k{=}100$ attains the L32 peak for DPG-Probe and DPG-Anchor.

k sensitivity concentrates at intermediate layers. The widest full-range k-bands appear around L0–L24 on semantic-OOD (max ≈ 3.3 pp for DPG-Probe at L0), where features are still transitioning



(a) Semantic-OOD (b) In-the-wild

Figure 8: **Neighborhood size k ablation — method comparison.** DPG variants are evaluated over $k \in \{5, \ldots, 15, 20, 25, 50, 100\}$ on the full SD v1.4 train reference bank. Shaded bands show the spread across k values at each layer. The dashed baseline uses frozen MetaCLIP-2 [CLS] at k=10.

Table 20: **k-ablation summary.** Peak-layer mean accuracy and mean k-band width (full range across k, averaged over all layers). Our default k=100 is highlighted.

Domain	Method	Peak Acc (%)	Mean k-band
Semantic-OOD (L36)	DPG-Probe	**95.7**	1.82 pp
	DPG-Anchor	95.6	1.92 pp
	DPG-AbsMean	94.1	1.73 pp
In-the-Wild (L32)	DPG-Probe	84.9	2.25 pp
	DPG-Anchor	84.4	2.28 pp
	DPG-AbsMean	**85.4**	2.78 pp
Default k=100 at peak layer (DPG-Probe):			
Semantic-OOD	DPG-Probe	**95.7**	0.78 pp
In-the-Wild	DPG-Probe	**84.9**	3.62 pp

from low-level to semantic structure. By L28–L36 the bands collapse to <0.45 pp, indicating that deep CLS-gradient embeddings induce a stable neighborhood structure regardless of k.

Method ranking is stable across k. DPG-Probe achieves the highest semantic-OOD peak (**95.7%** at L36, k=100; +11.7 pp over the [CLS] baseline at L36), while DPG-AbsMean marginally leads in-the-wild at k=100 (**85.4%** vs. **84.9%** for DPG-Probe). All three DPG variants dominate the frozen [CLS] baseline from L20 onward on both benchmark groups. Notably, DPG-AbsMean degrades at L48 on semantic-OOD (83.2% at k=100 vs. 93.7% at L36), whereas DPG-Probe and DPG-Anchor remain above 94%; This reinforces L32/L36 as the operating layers in the main experiments.

D.3 Summary and Implications

The k ablation supports the hyper-parameter choices in the main experiments:

- k=100 **is well justified.** At operating layers it equals the best k for DPG-Probe on both domain groups and lies within 0.4 pp of the semantic-OOD peak for DPG-AbsMean.
- **Layer selection matters more than k.** k-sensitivity is largest at shallow layers (up to ≈3.3 pp at L0) but negligible at L32/L36 ($\leq$0.8 pp band on semantic-OOD), where DPG-Probe features are semantically organized.
- **DPG-Probe offers the best overall trade-off.** It leads on semantic-OOD, stays within 0.5 pp of the in-the-wild leader, avoids the L48 DPG-AbsMean failure mode, and achieves the lowest combined k-variance at operating layers among the three aggregations.

E Robustness to Image Corruptions

We evaluate DPG-Probe alongside three supervised forensic detectors—DDA, OMAT, and DGSNet—under four common image corruption types at three severity levels. The goal is to assess whether the reference-based kNN design of DPG-Probe degrades gracefully when test images are corrupted, given that the reference bank remains clean throughout.

E.1 Experimental Setup

Corruption protocol. We apply four corruption types, each at three severity levels, to test images only; the DPG-Probe reference bank remains *clean* throughout:

- **JPEG compression:** quality factors $q \in \{90, 75, 50\}$;
- **Gaussian noise:** standard deviations $\sigma \in \{0.02, 0.05, 0.10\}$;
- **Gaussian blur:** standard deviations $\sigma \in \{1.0, 2.0, 3.0\}$;
- **Brightness darkening:** multiplicative factors $f \in \{0.8, 0.6, 0.4\}$.

Severity levels 1–3 correspond to mild, moderate, and strong degradation, respectively.

Datasets. We evaluate on the same 12 shared datasets used in the main experiments: six in-the-wild targets (WildRF, CO-SPYBench, Chameleon, AIGIBench, `picsum_images`, `midjourney_images`) and six semantic out-of-domain targets (Genimage++, AIGI-Holmes, AIGI-Now-pixel, AIGI-Now-semantic, anime-test, Community-Forensics-eval). Comparison baselines (DDA, OMAT, DGSNet) are also evaluated on these 12 datasets; results from any additional datasets in their benchmark are excluded to ensure a fair comparison.

Metrics. For DPG-Probe we report kNN accuracy (k=100, seed 1314, layer 32). For the supervised baselines we report standard classifier accuracy. Both reflect the fraction of correctly classified images and are directly comparable on the same test splits.

E.2 Results

Overview. Figure 9 shows the full 2×4 comparison across both domain groups and all four corruption types. DPG-Probe (kNN) maintains the highest mean accuracy curve across nearly all conditions, leading in 12 of 12 in-the-wild corruption-severity columns and all semantic-OOD columns. Table 23 summarises the grand mean across all 12 datasets and three severity levels: DPG-Probe achieves 85.8%, outperforming DDA (76.1%, $+9.7$ pp), OMAT (62.8%, $+23.0$ pp), and DGSNet (49.7%, $+36.1$ pp).

Per-corruption analysis. Across all four corruptions, DPG-Probe leads:

- **JPEG compression** (85.9%) vs. DDA (75.7%, $+10.2$ pp): Compression is the corruption type where DDA degrades most steeply (from 82.4% at sev. 1 to 76.3% at sev. 3 on in-the-wild), while DPG-Probe remains stable ($87.9\%\rightarrow86.8\%$).
- **Gaussian noise** (84.5%) vs. DDA (71.0%, $+13.5$ pp): DPG-Probe holds above 86% through sev. 2 across in-the-wild targets, whereas DDA drops from 76.1% to 69.8%. Notably, OMAT improves slightly at higher noise severities on semantic-OOD targets (peaking at σ=0.10, 77.5%), possibly because its training distribution included noisy images.
- **Gaussian blur** (83.0%) vs. DDA (74.8%, $+8.2$ pp): Blur is the hardest corruption for all methods. DPG-Probe degrades from 85.5% (sev. 1) to 80.2% (sev. 3) overall, while DDA drops from 81.8% to 67.7%.
- **Brightness darkening** (89.5%) vs. DDA (82.9%, $+6.6$ pp): This is the corruption type where DPG-Probe is most robust, retaining above 90% on semantic-OOD targets even at the strongest darkening (f=0.4). The advantage narrows slightly on in-the-wild targets at sev. 3 (81.9% vs. 80.5% for DDA), suggesting that strong darkening makes some in-the-wild fakes harder to distinguish even with geometry gradients.

In-the-Wild vs. Semantic-OOD. Figures 10 and 11 present the per-domain degradation curves. On semantic-OOD targets (Table 22), the gap between DPG-Probe and supervised baselines is generally larger because the baselines were not trained on these generator families, while DPG-Probe relies on a generator-agnostic geometry-gradient representation. On in-the-wild targets (Table 21), DDA is a stronger competitor, benefiting from diverse training data, but DPG-Probe still leads in 11 out of 12 corruption-severity columns.

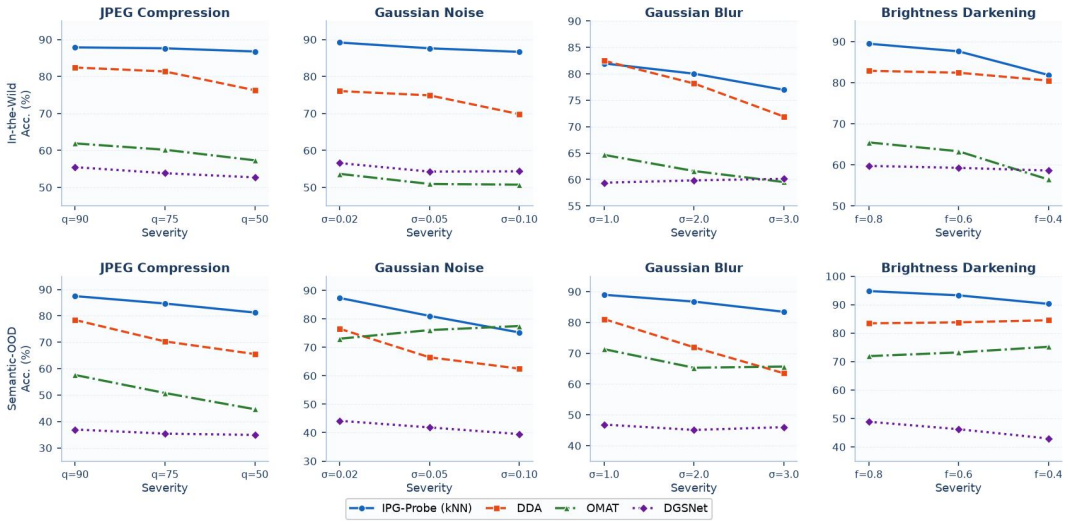


Figure 9: **Robustness comparison overview.** Mean accuracy (%) of DPG-Probe (kNN), DDA, OMAT, and DGSNet as a function of corruption severity (level 1–3) across four corruption types (columns) and two domain groups (rows). All values are averaged over the shared 12 datasets within each domain. Higher is better; DPG-Probe consistently dominates across all conditions.

Dataset-level breakdown. Figure 14 provides a per-dataset heatmap of accuracy averaged over the three severity levels. DPG-Probe achieves consistently high values (green) across both domain groups, while DGSNet shows widespread low values (red) and OMAT is highly inconsistent, performing well on some datasets (e.g., `midjourney_images`, Community-Forensics) but poorly on others (e.g., `anime_test`, AIGI-Now-semantic).

E.3 Discussion

These results demonstrate two structural advantages of DPG-Probe under corruption:

1. **Clean reference bank.** Because only the query image is corrupted while the reference bank remains clean, the kNN classifier can still draw on high-quality geometry-gradient features for nearest-neighbour comparison. This is in contrast to supervised classifiers that may have learnt shallow frequency or texture statistics which are directly destroyed by corruption.
2. **Geometry gradients are corruption-resistant.** The rotation self-consistency loss operates on mid-layer CLS representations (layer 32 of MetaCLIP-2), which are more semantically abstract than low-level pixel statistics. Corruptions that primarily alter low-frequency brightness or mid-frequency JPEG artefacts are less likely to change the relative geometry-inconsistency between real and generated images at this level of representation.

Taken together, the robustness evaluation confirms that DPG-Probe is not only more accurate under clean conditions but also more stable under realistic post-processing corruptions, with a mean gain of $+9.7$ pp over the best-performing supervised baseline (DDA) averaged across all corruption types and severity levels.

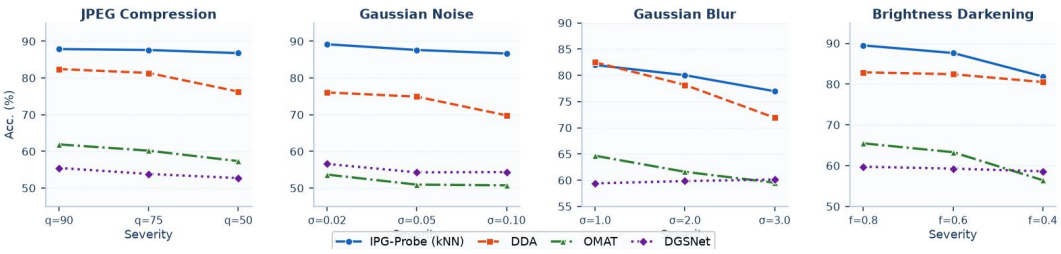


Figure 10: **In-the-Wild robustness.** Per-corruption degradation curves (mean over 6 in-the-wild datasets).

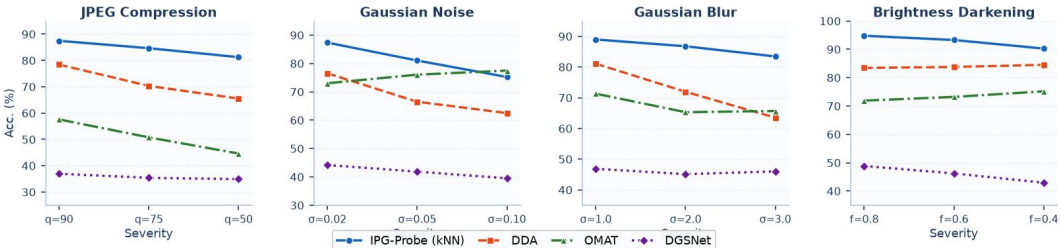


Figure 11: **Semantic-OOD robustness.** Per-corruption degradation curves (mean over 6 semantic out-of-domain datasets).

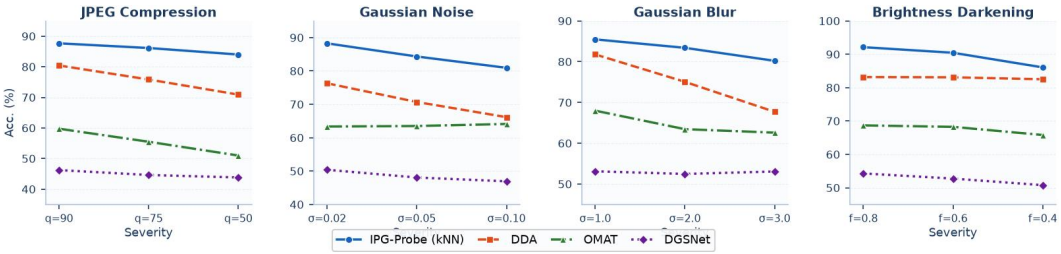


Figure 12: **Overall robustness degradation.** Mean accuracy across all 12 shared datasets at each severity level. DPG-Probe shows the mildest degradation slope across all four corruption types.

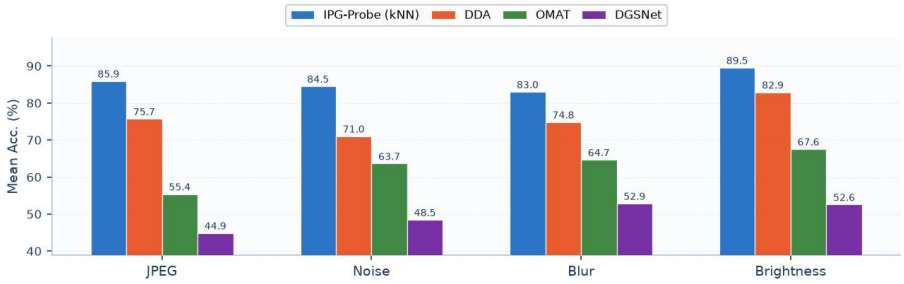


Figure 13: **Mean accuracy by method and corruption type.** Each bar averages over all 12 shared datasets and three severity levels. Numbers above bars indicate mean accuracy (%).

Table 21: **Robustness on In-the-Wild datasets.** Mean accuracy (%) over shared datasets at three corruption severity levels. Best result per column shown in **bold**. DPG-Probe uses kNN (k=100) with a clean reference bank; corruptions are applied to query images only.

Method	JPEG Compression			Gaussian Noise			Gaussian Blur			Brightness Darkening		
	q=90	q=75	q=50	σ=0.02	σ=0.05	σ=0.10	σ=1.0	σ=2.0	σ=3.0	f=0.8	f=0.6	f=0.4
DPG-Probe (kNN)	**87.9**	**87.6**	**86.8**	**89.2**	**87.6**	**86.7**	82.0	**80.0**	**77.0**	**89.5**	**87.6**	**81.9**
DDA	82.4	81.4	76.3	76.1	74.9	69.8	**82.5**	78.2	71.9	82.9	82.4	80.5
OMAT	61.9	60.2	57.3	53.7	50.9	50.7	64.7	61.6	59.5	65.5	63.3	56.4
DGSNet	55.5	53.8	52.7	56.6	54.3	54.4	59.4	59.8	60.1	59.7	59.3	58.6

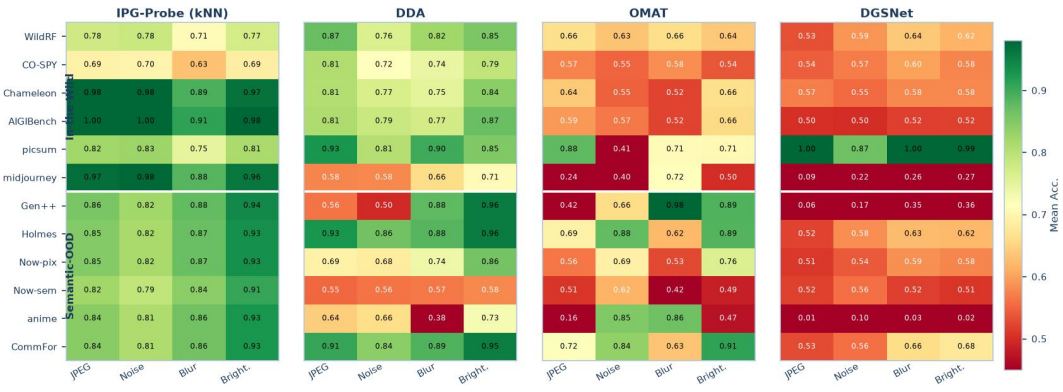


Figure 14: **Per-dataset accuracy heatmap (severity-averaged).** Each cell shows the mean accuracy across severity levels 1–3 for a given (dataset, corruption) pair. The horizontal white line separates In-the-Wild (top) from Semantic-OOD (bottom) datasets. DPG-Probe (leftmost panel) shows consistently high values across both groups.

Table 22: **Robustness on Semantic-OOD datasets.** Mean accuracy (%) over shared datasets at three corruption severity levels. Best result per column shown in **bold**. DPG-Probe uses kNN (k=100) with a clean reference bank; corruptions are applied to query images only.

Method	JPEG Compression			Gaussian Noise			Gaussian Blur			Brightness Darkening		
	q=90	q=75	q=50	σ=0.02	σ=0.05	σ=0.10	σ=1.0	σ=2.0	σ=3.0	f=0.8	f=0.6	f=0.4
DPG-Probe (kNN)	**87.5**	**84.7**	**81.2**	**87.4**	**81.0**	75.2	**89.0**	**86.8**	**83.4**	**94.8**	**93.3**	**90.3**
DDA	78.4	70.3	65.5	76.5	66.5	62.5	81.1	71.9	63.5	83.4	83.8	84.5
OMAT	57.6	50.8	44.6	73.0	76.1	**77.5**	71.3	65.3	65.7	71.9	73.2	75.2
DGSNet	37.0	35.4	34.9	44.2	41.8	39.5	46.9	45.1	46.0	48.8	46.2	42.9

Table 23: **Robustness summary.** Mean accuracy (%) across all 12 shared datasets and 3 severity levels. *Overall*: grand mean across all corruptions. Best result per column in **bold**.

Method	JPEG	Gaussian Noise	Gaussian Blur	Brightness Darkening	Overall
DPG-Probe (kNN)	**85.9**	**84.5**	**83.0**	**89.5**	**85.8**
DDA	75.7	71.0	74.8	82.9	76.1
OMAT	55.4	63.7	64.7	67.6	62.8
DGSNet	44.9	48.5	52.9	52.6	49.7