GLOBALFORGE: TOWARDS ROBUST AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review



Figure 1: Existing detectors read out local details (red box: an enlarged patch; heat map: Grad-CAM), whereas GlobalForge reads out global structure (green box: a cross-image region). On clean images both decide correctly. After common propagation degradations (JPEG, blur, resize), the local cue and its attention map dissolve and the prediction flips, while the global cue and its map persist; the curves sketch accuracy against degradation strength.

ABSTRACT

AI-generated image (AIGI) detectors achieve strong accuracy on clean benchmarks, but their performance drops sharply after images are propagated through real-world channels. We trace this fragility to what these detectors actually learn: they overfit to local artifacts left by generators in small spatial neighborhoods, which are easily destroyed by common propagation degradations such as JPEG compression and blur. Instead, we shift the discriminative cue from fragile local artifacts to more robust global structure. Building on this, we propose GlobalForge, a framework with two complementary modules. The Local Information Bottleneck (LIB) suppresses local components to block shortcut learning, while the Global Structural Reasoning (GSR) module forces every token to gather evidence from distant regions. Both modules are trained jointly under a contrastive structural loss based on degradation that keeps the resulting features stable under degradation. To support fine-grained robustness evaluation, we further introduce RealDeg-Bench, covering 7 common degradation operators and multi-step compound chains. GlobalForge improves average BAcc on 8 in-the-wild benchmark groups by $\mathbf{5.89}\%$ over the previous state-of-the-art, and is clearly ahead of representative baselines on RealDeg-Bench under both single and compound degradations. Code is available at `https://anonymous.4open.science/r/GlobalForge-BE0F/`.

1 INTRODUCTION

Recent generative models can synthesize images that are visually indistinguishable from real photographs. This has raised serious concerns about deepfakes, disinformation, and copyright infringement (Mirsky & Lee, 2021; Tolosana et al., 2020), and reliable detection of AI-generated images (AIGIs) has become an urgent need. Many detection methods have been proposed, and most achieve strong accuracy on clean benchmarks. But real deployment is far harsher. Images shared on social media are repeatedly compressed, color-adjusted and re-edited. These operations build up into

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

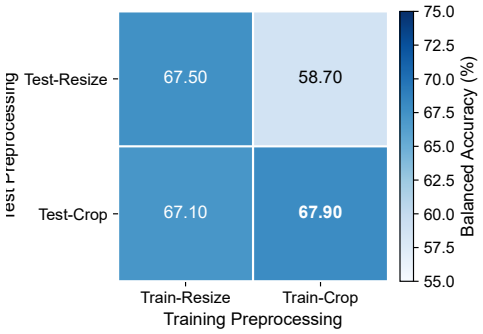

Figure 2: **Preprocessing cross-experiment (UnivFD on GenImage).** Rows: test-time preprocessing; columns: training-time. Crop-trained models suffer a sharp accuracy drop when evaluated with resize, revealing reliance on spatially sensitive local cues.

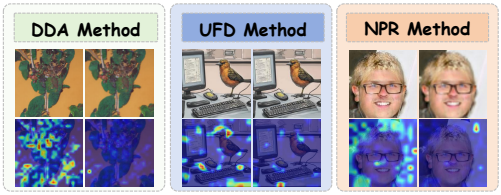

Figure 3: **Attention collapse under image degradation.** Grad-CAM visualizations for DDA, UnivFD, and NPR under different degradations. For each method, the top row shows the inputs and the bottom row shows the corresponding attention maps; within each method, the left column is the clean image and the right column is its degraded counterpart. The three methods are evaluated under resizing, JPEG compression, and blur, respectively.

compound degradation chains (Corvi et al., 2023b; Cozzolino et al., 2024), making current detectors lose substantial accuracy (Pal et al., 2024).

The drops are not random failures. They reveal that the cue these detectors learn from is the wrong one. Such detectors overfit to local artifacts left by generators in small spatial neighborhoods, and any propagation degradation can destroy these artifacts directly. We give a detailed diagnosis in Section 3. Most prior work still treats local artifacts as the right signal to defend. Data augmentation (Yan et al., 2024a; Pal et al., 2024) broadens the training distribution but remains a passive defense, and prior work (Pal et al., 2024) shows it generalizes poorly to unseen degradations. Image restoration (Ke & Wang, 2023) tries to recover damaged artifacts, but they already lie outside the natural-image distribution and restoration pulls the image back toward natural-image statistics. Neither line addresses the problem at its root.

We argue for a different path. Instead of defending fragile local artifacts, we intervene in what the model is allowed to learn from, and steer it toward global structural coherence, by which we mean cues that depend on long-range relationships across the image rather than on any single local neighborhood. Such cues are not easy to disrupt with local perturbations, and they do not depend on the local signature of any specific generator, so they help with degradation robustness and cross-generator generalization at the same time.

Following this idea, we propose GlobalForge, a detection framework with two complementary modules. A Local Information Bottleneck (LIB) suppresses the fragile local components on which shortcut learning would otherwise latch. A Global Structural Reasoning (GSR) module forces every token to gather evidence from distant regions, so the model has to rely on long-range structure for the decision. The two modules are trained jointly under a degradation-aware contrastive structural loss ($\mathcal{L}_{\text{DCS}}$) that further keeps the resulting global representation stable under realistic degradation. Our contributions are summarized as:

- We identify local-artifact dependency as the root cause of degradation fragility in existing AIGI detectors, and a controlled experiment shows that blocking the model's access to local artifacts improves both degradation robustness and cross-generator generalization.
- We propose **GlobalForge**, to our knowledge, the first AIGI detection framework that actively suppresses local shortcuts and forces the discriminative decision onto long-range structural cues, achieving a robustness profile that is decoupled from generator-specific local signatures.
- We introduce **RealDeg-Bench**, a robustness-oriented evaluation benchmark for fine-grained diagnosis of AIGI detectors under common single-operator degradations and multi-step compound chains that approximate real social-media propagation.
- GlobalForge improves average BAcc on 8 in-the-wild benchmark groups by $\mathbf{5.89}\%$ over the previous state-of-the-art, and reaches $\mathbf{85.81}\%$ average BAcc on RealDeg-Bench, clearly ahead of representative baselines under both single and compound degradations.

2 Related Work

2.1 AI-Generated Image Detection

Existing AIGI detectors fall into three paradigms. **Pretrained-representation-based** detectors freeze or lightly adapt the features of large pretrained models such as CLIP, and treat real/fake detection as a binary classification task on top of these general-purpose representations (Wang et al., 2020; Ojha et al., 2023; Liu et al., 2024; Tan et al., 2025; Yan et al., 2024a; Chen et al., 2025; Qin et al., 2025). **Pixel or frequency statistics based** detectors target low-level generative fingerprints such as spectral artifacts (Frank et al., 2020), upsampling grid traces (Tan et al., 2024), frequency inconsistency (Li et al., 2025b), and diffusion reconstruction error (Wang et al., 2023; Chen et al., 2024). **Semantic and structure based** methods such as LASTED (Wu et al., 2025) and AntifakePrompt (Chang et al., 2023) instead use higher-level signals, and share the global-consistency spirit advocated in this work. Note that reading out a global token does not by itself imply relying on global structure: nothing constrains where a CLS token's evidence comes from, whereas GlobalForge constrains it explicitly.

A growing line of work directly targets robustness under propagation degradations. AIDE (Yan et al., 2024a) and Semi-Truths (Pal et al., 2024) broaden training distributions through aggressive degradation augmentation, but Semi-Truths itself reports that augmentation alone does not shift detectors' feature preference. B-Free (Guillaro et al., 2025) addresses platform-induced bias through a debiased training paradigm, while DF-UDetector (Ke & Wang, 2023) restores degraded inputs to recover artifacts, yet the restored statistics drift toward natural images rather than the abnormal traces of generators. TextureCrop (Konstantinidou et al., 2025) and Karageorgiou et al. (2025) go the opposite way and reinforce the local-artifact cue, which helps on clean benchmarks but widens the gap once the image is degraded. These efforts remain largely passive: they patch robustness on the input or training side without changing the cue the detector relies on, so accuracy still degrades sharply along realistic propagation chains.

2.2 AIGI Detection Evaluation Benchmarks

Early AIGI benchmarks mostly target cross-generator generalization. ForenSynths (Wang et al., 2020) evaluates detectors on GANs, and GenImage (Zhu et al., 2023) extends this to million-scale diffusion images across multiple generators, becoming the de facto training and evaluation suite for modern detectors. Chameleon (Yan et al., 2024a) further tracks detector performance against rapidly evolving generators. More recent efforts move toward deployment-oriented evaluation. WildRF (Cavia et al., 2024) and BFree-Online (Guillaro et al., 2025) collect images directly from social-media platforms to approximate in-the-wild conditions, while real-chain (Liu et al., 2025) captures images that have already traversed multi-step social-platform degradations.

These benchmarks have two structural limitations for studying degradation robustness. First, their robustness protocols typically apply only one perturbation at a time, such as JPEG at a fixed quality factor or Gaussian blur at a single strength (Corvi et al., 2023b; Cozzolino et al., 2024), which makes it hard to characterize how accuracy decays along realistic degradation chains. Second, in-the-wild benchmarks rely on degradations that arise incidentally from platform processing, with uncontrollable type, strength, and depth, so they cannot pinpoint which operators or compositions drive the failures.

3 Motivation and Design Principles

As discussed in Section 1, AIGI detectors that perform near-saturation on clean benchmarks routinely collapse once images traverse real propagation chains. We argue that this fragility is a structural property of what existing detectors learn. Generators leave behind local artifacts in small spatial neighborhoods (Frank et al., 2020; Corvi et al., 2023a; Tan et al., 2024), and these artifacts all sit in the local part of the image. Common propagation degradations such as JPEG quantization, resize resampling, and blur smoothing all perturb the local spectrum. They therefore destroy these artifacts with little effort. The artifacts are also generator-specific, since different generators apply different local operators, so detectors trained on them struggle on unseen generators. They behave less like useful cues and more like shortcuts (Geirhos et al., 2018; 2020) that crowd out more reliable evidence.

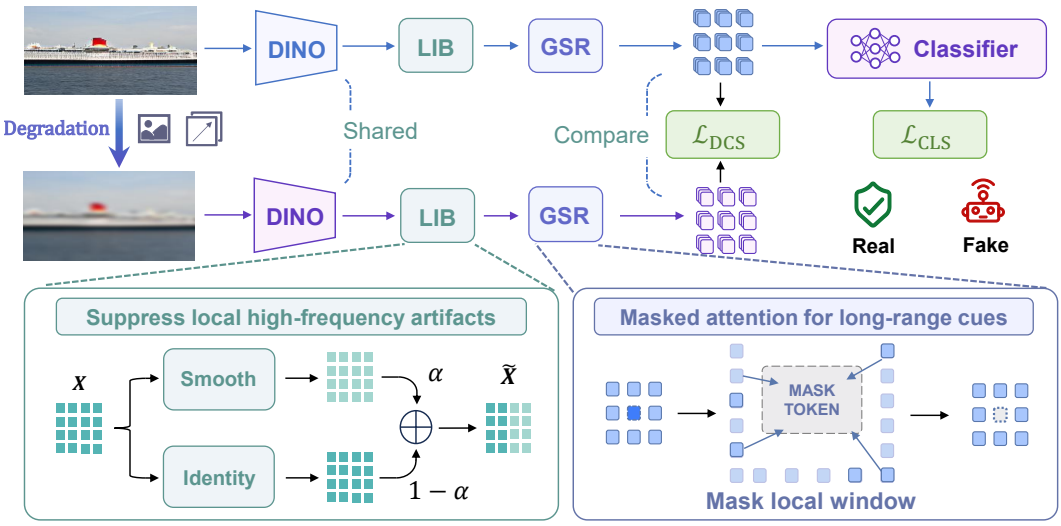

Figure 4: Overall architecture of GlobalForge. Clean and degraded views share a DINO backbone. LIB suppresses local high-frequency features through a learnable gate α; GSR masks each query's local window so that tokens aggregate only distant evidence; $\mathcal{L}_{\text{DCS}}$ aligns the two views and $\mathcal{L}_{\text{cls}}$ supervises the prediction.

Common practice in the field already points in this direction. Existing AIGI detectors almost universally adopt center crop rather than resize at inference time (Wang et al., 2020; Ojha et al., 2023; Tan et al., 2024), because resize causes a substantial drop in detection accuracy. Yet resize preserves both the semantic content and the spatial layout of an image, so the cue it disrupts cannot be semantic or structural in nature. A preprocessing cross-experiment that varies training and test time preprocessing independently confirms this asymmetry directly (Fig. 2). A model trained with center crop loses about **8.8%** of BAcc when tested with resize, while a model trained with resize barely degrades when tested with resize at all. The cue that collapses must therefore be a fine-grained signal sensitive to the spatial arrangement of pixel neighborhoods, in agreement with the local-artifact characterization above.

The same fragility extends from resize to other common degradations and shows up directly in the model's attention. Grad-CAM visualizations of three representative detectors (Fig. 3) show that on clean images the attention of all three concentrates sharply on local patches, and that a mild degradation dissolves these focal regions into diffuse, noise-like maps. We treat these maps as qualitative corroboration; the quantitative test is the intervention in Section 6.4, where removing the constraint on local evidence costs 6.1% BAcc.

This diagnosis motivates GlobalForge's two core design principles: suppress local information through LIB and explicitly guide global structural reasoning through GSR.

4 Method

The overall architecture of GlobalForge is shown in Figure 4. Given an input image $I \in \mathbb{R}^{3\times H_0\times W_0}$, our goal is to learn a discriminative mapping $f_\theta : I \to y \in \{0, 1\}$ that remains reliable under real-world degradation. The backbone features first pass through **LIB** and then **GSR**, and are read out by a classification head. The two modules are trained jointly under a degradation-aware contrastive structural loss ($\mathcal{L}_{\text{DCS}}$) that aligns clean and degraded views in feature space. The components are detailed in Sections 4.1–4.3.

4.1 Local Information Bottleneck (LIB)

To prevent the model from overfitting to fragile local artifacts, LIB introduces a learnable "original-feature / smoothed-feature" fusion pathway on the deep features of the backbone, thereby suppressing local high-frequency components in a controlled manner.

A natural alternative is low-pass filtering in the pixel or frequency domain of the input. We instead operate in the feature domain: input-side filtering leaves visible side effects such as ringing near edges, which the model can latch onto as new shortcuts and which themselves shift under degradation, whereas suppression after encoding shapes *what the classifier sees*, not the raw input distribution (compared in Section 6.4).

Specifically, let the patch tokens produced by the backbone be $\mathbf{F} \in \mathbb{R}^{B\times N\times d}$, and reshape them into a spatial feature map $\mathbf{X} \in \mathbb{R}^{B\times d\times h\times w}$, where B is the batch size, $N = h \times w$ is the number of patch tokens, d is the channel dimension, and h, w are the spatial height and width of the feature map. We construct a fixed $k \times k$ Gaussian kernel (with $\sigma = k/3$) and obtain a smoothed version $\mathbf{X}_{\text{smooth}}$ via depthwise convolution. We introduce a single learnable scalar β and map it through a sigmoid to a fusion coefficient $\alpha = \sigma(\beta) \in (0, 1)$. The final output is the convex combination of the two:

$$\hat{\mathbf{X}} = (1 - \alpha)\mathbf{X} + \alpha\mathbf{X}_{\text{smooth}}. \tag{1}$$

Gaussian smoothing acts as a low-pass filter, attenuating the high-frequency feature components where local artifacts concentrate (upsampling grids, noise residues) while preserving the lower-frequency structural information. As $\alpha \to 0$ the output degenerates to the original feature; the underlying raw gate scalar β (from which $\alpha = \sigma(\beta)$ is derived) is optimized jointly with the rest of the model.

4.2 Global Structural Reasoning (GSR)

LIB suppresses local high-frequency components along the frequency axis, but the model may still exploit local correlations between spatially adjacent tokens. GSR closes this loophole by forcing each token to gather evidence exclusively from distant regions, complementing LIB along the spatial axis.

Specifically, the LIB output $\hat{\mathbf{X}} \in \mathbb{R}^{B\times d\times h\times w}$ is reshaped back to a token sequence $\hat{\mathbf{F}} \in \mathbb{R}^{B\times N\times d}$. GSR first computes the standard self-attention score matrix:

$$\mathbf{S} = \frac{(\hat{\mathbf{F}}\mathbf{W}_q)(\hat{\mathbf{F}}\mathbf{W}_k)^\top}{\sqrt{d}}. \tag{2}$$

To block the model from exploiting neighborhood local correlations, we apply a local masking window at each position before the Softmax: all attention entries whose Chebyshev distance (i.e., chessboard distance $d_{\text{Cheb}}(u, v) = \max(|u_x - v_x|, |u_y - v_y|)$) to the query position is at most w_{gsr} (default $w_{\text{gsr}} = 3$) are set to $-\infty$:

$$\tilde{\mathbf{S}}_{u,v} = \begin{cases} -\infty, & \text{if } d_{\text{Cheb}}(u, v) \le w_{\text{gsr}}, \\ \mathbf{S}_{u,v}, & \text{otherwise.} \end{cases} \tag{3}$$

The masked scores are then normalized by Softmax, used to update the features, and passed through an output projection with a residual connection:

$$\mathbf{F}_{\text{GSR}} = \text{Proj}\big(\text{Softmax}(\tilde{\mathbf{S}}) \cdot \hat{\mathbf{F}}\mathbf{W}_v\big) + \hat{\mathbf{F}}. \tag{4}$$

During training, masking is applied with probability $p_{\text{mask}} = 1.0$ by default, with sensitivity to the masking probability and window size analyzed in Section 6.4. This design forces each token's representation update to come exclusively from distant regions, so the model can only discriminate by integrating cross-region structural information. Combined with the frequency-level suppression of LIB, this steers the model toward a global structural representation that is more stable under both degradation and generator shifts. Although backbone tokens have already attended globally, their readout remains locally anchored: if it were not, masking the local window would be inert, yet removing the mask costs 6.1% BAcc (Table 4).

4.3 Synergistic Optimization Objectives

The overall training loss of GlobalForge consists of a classification term and a degradation-aware contrastive structural term:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{dcs}}\mathcal{L}_{\text{DCS}}, \tag{5}$$

where $\mathcal{L}_{\text{cls}}$ is the binary cross-entropy loss with label smoothing, $\mathcal{L}_{\text{DCS}}$ is an auxiliary contrastive regularizer (defined below), and $\lambda_{\text{dcs}} = 0.01$ by default. $\mathcal{L}_{\text{DCS}}$ serves as an auxiliary regularizer that explicitly aligns paired clean and degraded views at the representation level. For each input I_i, we online-generate a compound-degraded view $\tilde{I}_i$ (random JPEG compression, Gaussian blur, and color perturbation), and denote the global image representations of the clean and degraded views as $\mathbf{z}_i$ and $\tilde{\mathbf{z}}_i$. $\mathcal{L}_{\text{DCS}}$ then takes the symmetric InfoNCE (Oord et al., 2018) form:

$$\mathcal{L}_{\text{DCS}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(\text{sim}(\mathbf{z}_i,\tilde{\mathbf{z}}_i)/\kappa)}{\sum_j \exp(\text{sim}(\mathbf{z}_i,\tilde{\mathbf{z}}_j)/\kappa)} + \log\frac{\exp(\text{sim}(\tilde{\mathbf{z}}_i,\mathbf{z}_i)/\kappa)}{\sum_j \exp(\text{sim}(\tilde{\mathbf{z}}_i,\mathbf{z}_j)/\kappa)}\right], \tag{6}$$

where $\text{sim}(\cdot,\cdot)$ is cosine similarity and κ is a temperature coefficient. This objective encourages the model to preserve degradation-invariant structural information rather than letting degradation signals leak into the final embedding.

5 REALDEG-BENCH: A COMPOUND-DEGRADATION ROBUSTNESS BENCHMARK

Motivation. Existing AIGI robustness benchmarks either test one perturbation at a time or rely on in-the-wild images with uncontrolled degradation types, strengths, and histories. However, real image propagation often involves compound degradation chains, such as re-compression, resizing, and re-uploading. To better characterize robustness under such realistic propagation processes, we introduce **RealDeg-Bench**, a compound-degradation benchmark that evaluates detector performance across controllable degradation depths and strengths.

Construction. RealDeg-Bench is built on T2I-CoReBench (Li et al., 2025c), which covers mainstream generators across GANs, diffusion models, and visual autoregressive architectures. We define a pool of seven degradation operators commonly observed in real propagation: JPEG compression, Gaussian blur, resize, Gaussian noise, brightness, contrast, and saturation, to cover real-world propagation conditions (full lists in Appendix A).

Evaluation protocol. RealDeg-Bench provides two complementary regimes. *Single-operator* conditions apply one operator per image, while strength-sweep curves at each fixed level are also reported (Figure 5). *Compound* conditions chain $N \in \{1,\ldots,5\}$ operators sampled independently and uniformly from the seven-operator pool *with replacement*, with each operator's strength independently drawn from its strength set:

$$\tilde{I} = \mathcal{D}_{k_n}(\cdots\mathcal{D}_{k_1}(I;\theta_1)\cdots;\theta_n), \quad k_i \sim \text{Uniform}(\mathcal{P}),\ \theta_i \sim \text{Uniform}(\mathcal{S}_{k_i}), \tag{7}$$

where $\mathcal{P}$ is the operator pool and $\mathcal{S}_{k_i}$ is the strength set of operator k_i. With-replacement sampling allows the same operator to appear multiple times in a chain at different strengths, reflecting realistic propagation where an image is repeatedly re-compressed, re-resized, or re-color-shifted as it traverses successive platforms. In total, RealDeg-Bench contains 13 conditions (1 clean, 7 single-operator, 5 compound) with 95,589 images. Construction details, exact strength sets, and dataset statistics are provided in Appendix A.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate on two complementary suites. *(1) Robustness suite.* RealDeg-Bench (Section 5), for controllable single and multi-step degradations. *(2) In-the-wild suite.* We evaluate on 8 benchmark groups: Chameleon (Yan et al., 2024a), SynthWildx (Cozzolino et al., 2024), WildRF (Cavia et al., 2024), AIGIBench (Li et al., 2025d), CO-SPY (Cheng et al., 2025), RR-Dataset (Li et al., 2025a), BFree-Online (Guillaro et al., 2025), and real-chain (Liu et al., 2025). These collectively cover platform-induced degradations, evolving generators, and naturally chained social-media propagation, with no additional degradation applied at evaluation time, so the results faithfully reflect open-distribution conditions. Results on six standard benchmarks, including UniversalFakeDetect (Ojha et al., 2023) and GenImage (Zhu et al., 2023), are reported in Appendix B.

Table 1: Full fine-grained results on **RealDeg-Bench** (BAcc %). Average is the mean of the three group means (Clean, Multi-Kind, Compound). Bold: best Average; underline: second best.

Degradation	NPR	UnivFD	C2P-CLIP	FatFormer	SAFE	AIDE	Effort	DRCT	Aligned	B-Free	DDA	GAPL	**Ours-d2**	**Ours-d3**
Clean	54.05	50.15	52.89	57.93	49.62	63.79	50.70	74.79	47.46	84.44	84.77	86.31	89.31	87.77
Multi-Kind Degradation														
Jpeg	54.10	50.12	52.82	56.46	49.53	63.92	50.71	74.75	47.40	84.40	84.84	86.19	89.35	87.79
Gaussian blur	53.73	47.31	56.50	60.39	50.82	50.48	50.31	64.12	51.37	76.95	79.09	81.59	69.95	84.65
Resize	53.96	48.49	51.60	64.92	50.83	55.69	50.93	69.55	53.99	73.89	80.17	85.92	87.08	89.63
Gaussian noise	53.56	46.86	52.65	46.10	49.37	57.92	52.91	73.92	41.64	81.56	70.53	75.37	70.37	82.06
Brightness	56.55	48.38	52.06	53.39	49.60	62.53	51.94	74.65	47.37	84.36	85.15	86.64	89.68	88.63
Contrast	55.94	48.71	55.45	49.81	50.31	58.60	52.76	73.75	47.76	83.97	84.51	83.18	89.53	87.69
Saturation	51.14	49.68	52.85	49.77	50.51	61.71	50.37	72.25	46.57	84.54	85.32	84.75	87.12	85.30
Compound Degradation														
Step1	54.63	48.41	53.48	54.56	50.05	58.86	51.68	71.96	48.36	81.80	81.40	83.89	83.54	86.87
Step2	54.65	46.75	53.60	54.32	50.33	56.23	52.51	69.44	48.10	78.60	78.07	79.87	79.34	85.18
Step3	54.55	47.06	53.00	54.71	50.24	55.82	52.58	66.32	48.44	76.59	74.52	76.14	74.58	82.81
Step4	55.69	45.98	53.36	54.23	50.08	56.73	51.97	64.08	48.52	74.49	71.16	72.38	72.14	81.27
Step5	55.26	46.56	52.59	53.35	50.24	55.34	52.44	62.28	47.62	72.94	68.59	71.27	69.24	79.53
Average	54.38	48.54	53.17	55.53	49.98	59.69	51.45	71.15	47.90	80.90	80.30	82.13	<u>82.79</u>	**85.81**

Evaluation Protocol. We adopt Balanced Accuracy (BAcc) as the primary metric, which provides a reliable and class-balanced assessment of detection performance. We compare GlobalForge with 12 representative detectors, covering both artifact-driven methods (Tan et al., 2024; Ojha et al., 2023; Liu et al., 2024; Li et al., 2025b; Yan et al., 2024a;b; Tan et al., 2025) and alignment-based approaches (Chen et al., 2024; Rajan et al., 2024; Guillaro et al., 2025; Chen et al., 2025; Qin et al., 2025). Six baselines (Tan et al., 2024; Ojha et al., 2023; Liu et al., 2024; Li et al., 2025b; Yan et al., 2024a;b) are retrained from scratch on the aligned dataset (Chen et al., 2025) to prevent detectors from latching onto format level shortcuts, while the remaining six (Tan et al., 2025; Chen et al., 2024; Rajan et al., 2024; Guillaro et al., 2025; Chen et al., 2025; Qin et al., 2025) are evaluated with their officially released weights, as their training is closely related to their own proposed datasets or specialized objectives and cannot be cleanly transferred to a unified pipeline without distorting the method. A detailed comparison between official and retrained weights, together with the rationale for this division, is provided in Appendix C.

Implementation Details. Following the experimental settings of DDA (Chen et al., 2025), we build GlobalForge on two pretrained ViT-L backbones from different families, DINOv2-L (Oquab et al., 2024) and DINOv3-L (Siméoni et al., 2025), and fine-tune them with LoRA (Hu et al., 2022). The training data, shared with the six retrained baselines, consists of real images paired with their VAE reconstructed counterparts as fake samples. We refer to these two variants as **Ours-d2** and **Ours-d3** throughout the experiments. Both variants are trained on the DDA-aligned dataset (Chen et al., 2025) described above. Full hyperparameters and training schedules are reported in Appendix D.

6.2 Controllable Degradation Evaluation on *RealDeg-Bench*

We evaluate all methods on RealDeg-Bench following the protocol in Section 5 and Appendix A.

Table 1 reports fine grained results. GlobalForge achieves the best or second best score under Clean, the vast majority of Multi-Kind Degradations, and *all* five Compound Degradation levels, with an average BAcc of **85.81%**. As shown in Figure 5, GlobalForge maintains a more stable trend as the strength of each single operator grows, while GAPL, DDA, and B-Free exhibit pronounced step-like drops, reflecting the fragility of local-texture cues under heavier perturbations. The advantage becomes more pronounced under Compound Degradation, where baselines suffer cliff-like drops as each additional operator destroys the local artifacts they rely on, whereas GlobalForge decays smoothly from Step1 to Step5, confirming the stability of global structural features under sequential degradation.

6.3 In-the-Wild Comprehensive Evaluation

We evaluate GlobalForge and 12 baselines on 17 in-the-wild test subsets across the 8 benchmark groups of Section 6.1, without applying any additional degradation.

Table 2: Performance comparison on In-the-wild Benchmarks(BAcc %). The Avg B.Acc is computed by first averaging within each of the 8 benchmark groups and then averaging the 8 group means. Bold indicates the best Avg B.Acc; underline indicates the second-best Avg B.Acc.

Method	Cham.	SynthWildx			WildRF			AIGIBench		CO-SPY					RR Dataset	BFree Online	Real-chain	Avg B.Acc
		DALLE3	Firefly	Midj.	FB	Reddit	Twitter	SocRF	ComAI	Civitai	DALLE3	instavibe	Lexica	Midj.v6				
AIDE (Yan et al., 2024a)	64.48	68.99	52.84	84.84	67.50	66.93	72.30	60.62	68.69	90.80	43.55	13.70	27.90	94.90	56.91	67.35	45.09	61.31
Aligned (Rajan et al., 2024)	61.30	49.60	53.90	52.40	48.40	54.00	40.60	51.00	60.20	7.50	8.40	6.20	11.50	9.70	47.70	38.10	49.13	45.02
DDA (Chen et al., 2025)	83.50	91.10	84.70	91.60	85.30	82.50	89.30	79.90	88.90	99.70	94.60	35.40	73.70	97.90	70.30	81.20	65.81	80.04
B-Free (Guillaro et al., 2025)	78.30	96.10	92.30	95.30	95.60	85.50	96.70	84.90	79.70	98.80	97.40	5.90	53.60	98.60	69.50	87.10	57.91	79.14
C2P-CLIP (Tan et al., 2025)	57.60	49.60	57.90	49.60	51.90	67.60	40.40	58.10	50.40	0.00	4.50	0.20	0.20	23.20	50.00	32.70	51.30	44.64
DRCT (Chen et al., 2024)	79.80	85.90	58.90	90.50	90.30	66.80	79.60	71.30	84.60	99.10	82.10	22.00	55.80	94.80	58.20	77.10	64.43	73.20
NPR (Tan et al., 2024)	62.73	64.09	56.30	62.35	63.75	60.40	56.92	60.18	64.26	72.90	71.78	62.80	77.20	71.60	57.24	55.95	45.55	59.53
SAFE (Li et al., 2025b)	51.80	53.62	52.96	53.61	50.31	53.73	53.08	52.52	54.81	99.70	98.13	95.40	98.70	98.30	56.76	55.19	63.09	60.54
UnivFD (Ojha et al., 2023)	74.46	53.04	57.12	55.72	46.88	50.07	53.05	52.80	79.55	55.90	82.62	66.50	96.40	61.00	49.41	69.68	62.27	62.47
GAPL (Qin et al., 2025)	69.28	87.22	87.49	84.71	88.13	82.13	88.92	80.10	75.96	99.00	93.64	94.70	21.90	98.20	78.12	68.40	63.45	76.45
Effort (Yan et al., 2024b)	69.03	55.20	56.54	55.67	45.31	51.13	48.18	50.72	73.88	99.10	97.76	87.20	99.10	85.10	52.61	59.30	65.36	63.28
FatFormer (Liu et al., 2024)	62.10	54.53	57.48	61.75	62.81	61.47	58.85	52.75	64.08	95.90	75.14	67.40	81.90	76.60	50.44	54.16	53.45	59.62
Ours-d2	88.05	93.78	85.25	92.20	86.19	87.73	95.80	82.68	93.19	99.60	99.50	55.70	93.70	99.30	74.22	79.99	65.77	<u>83.23</u>
Ours-d3	88.03	92.69	90.18	92.12	91.88	94.40	95.90	87.18	91.55	99.80	99.25	65.50	93.20	98.70	75.93	84.21	72.86	**85.93**

Table 3: Ablation of GlobalForge core components on **In-the-Wild Benchmarks** and **RealDeg-Bench** (BAcc %). Left: single-component activation; right: single-component removal.

only			w/o		
Config.	In-the-wild	RealDeg	Config.	In-the-wild	RealDeg
LIB only	81.23	81.47	w/o LIB	83.03	83.80
GSR only	82.31	82.74	w/o GSR	83.12	83.97
DCS only	80.98	80.24	w/o DCS	83.94	82.61
Full	**85.93**	**85.81**	Baseline	80.21	80.03

Table 4: Alternative analysis of input-side suppression and receptive-field variants on **RealDeg-1Step** (BAcc %).

Configuration	Components	BAcc
Data Aug	none	82.56
Blur (LIB→pix.)	GSR+DCS	81.39
FFT (LIB→freq.)	GSR+DCS	80.12
Global Attn. (GSR→glob.)	LIB+DCS	80.73
Full (Ours)	**LIB+GSR+DCS**	**86.87**

Table 2 shows that GlobalForge achieves the best average BAcc of **85.93%**, outperforming the strongest baseline DDA by 5.89%. It ranks first on 3 of 8 groups and remains highly competitive on the others, showing consistent robustness across platforms and generators. Notably, while some baselines collapse on out-of-distribution subsets, GlobalForge maintains strong performance under diverse real world propagation conditions.

6.4 Ablation and Mechanism Analysis

Component necessity. In Table 3, baseline is the plain fine-tuned backbone without any proposed component, trained with the exact same data, backbone and other settings listed in Appendix D as GlobalForge. Removing any of LIB, GSR, and DCS leads to a clear drop. The decomposition is informative: DCS alone adds only 0.8%/0.2% over the baseline, whereas LIB+GSR without DCS already recovers 3.7%/2.6% of the 5.7%/5.8% total, so robustness stems primarily from the structural inductive bias, with contrastive invariance complementary only once the cue has been shifted.

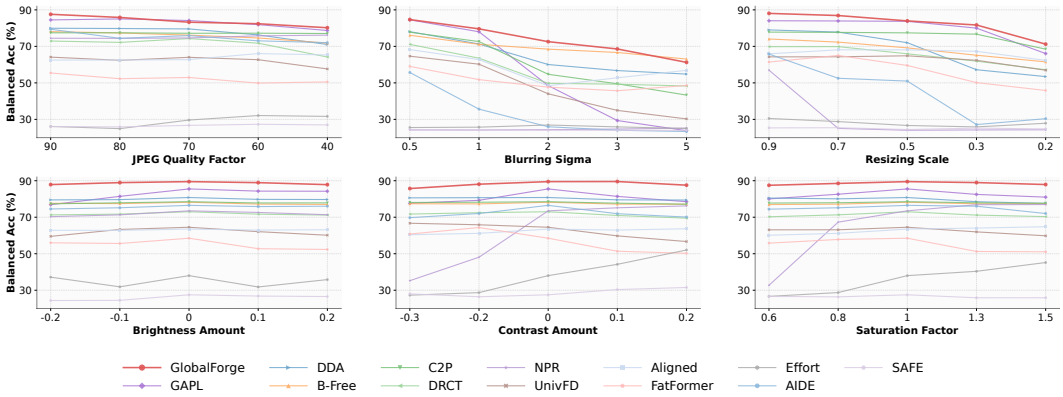

Figure 5: Fine-grained strength performance curves on single degradations on **RealDeg-Bench**. The horizontal axis denotes the strength level and the vertical axis denotes B.Acc (%).

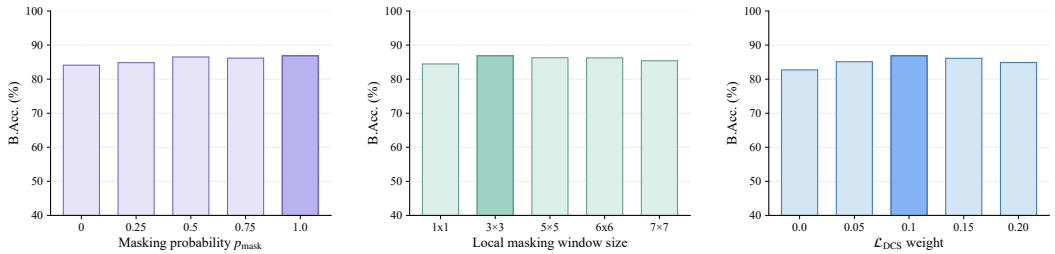

Figure 6: Sensitivity analysis of our method on RealDeg-Bench under the single degradation setting, showing the impact of the masking probability p_{mask}, the local masking window size, and the weight of $\mathcal{L}_{\text{DCS}}$.

Alternative analysis. Table 4 tests whether GlobalForge can be replaced by simpler alternatives. For fairness, each variant replaces only the target module while keeping all other components unchanged. Aug examines whether standard degradation augmentation alone is sufficient. Blur and FFT replace LIB with input-side suppression by pixel-domain smoothing and input-frequency low-pass filtering respectively. Consistent with the design motivation in Section 4, the results show that LIB is not merely input-side low-pass filtering. LIB operates after backbone encoding and is preferable to ordinary low-pass filtering, which can introduce new artifacts. The particularly low performance of FFT may be due to frequency artifacts coming from the operation itself, which the model may overfit as new shortcuts. Global Attention replaces GSR with unconstrained global access, showing that explicit local-window masking is needed to force long-range structural reasoning. If GlobalForge merely enforced low-pass filtering and perturbation invariance, Blur, FFT, and Aug would match it; they trail by 5.5%, 6.8%, and 4.3% BAcc, so the gain is not reducible to either ingredient. Detailed experimental settings are reported in Appendix D.

Sensitivity analysis. We analyze the sensitivity of our method to three hyperparameters covering both GSR and DCS: the masking probability p_{mask}, the local masking window size, and the weight of $\mathcal{L}_{\text{DCS}}$, evaluated under the single-step degradation setting of RealDeg-Bench (Figure 6). Accuracy peaks at $p_{\text{mask}} = 1.0$ for GSR, indicating that always forcing each token to aggregate evidence from outside its local window is the most effective. For the masking window size and the weight of $\mathcal{L}_{\text{DCS}}$, a moderate setting works best. The default configuration consistently achieves the best overall trade-off.

7 Conclusion and Discussion

From a feature preference perspective, we identify local artifact dependency as a key cause of robustness failures in AIGI detectors. Local artifacts are highly vulnerable to realistic degradations, so detectors relying on them remain intrinsically fragile. Common degrade then restore defenses offer only limited improvement. To address this, we propose GlobalForge, which actively suppresses such fragile local cues and guides the model toward global consistency, a more stable criterion for distinguishing real from fake images. With this shift in feature preference, GlobalForge improves both robustness to degradation and generalization to unseen generative paradigms. In addition, we introduce RealDeg-Bench, a reproducible robustness benchmark inspired by common social media propagation distortions.

AI Use Statement

In this work, we did not use generative AI tools for any task that requires disclosure. Generative AI tools were used only to improve the readability of the manuscript, namely to correct grammar, refine imprecise wording, and shorten some passages, and to help polish scientific figures. All AI-assisted content has been reviewed by the authors: every edited passage was compared against the original text and the corresponding results, and every AI-assisted figure was checked against the underlying data; no claims, numbers, or references were introduced by AI. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility Statement

We have made efforts to ensure that our results are reproducible. Anonymous source code and pretrained weights are available at the repository linked in the abstract, and the full RealDeg-Bench data and construction scripts will be released upon acceptance. The construction protocol of RealDeg-Bench, including the exact degradation operators, strength sets, sampling procedure, and dataset statistics, is described in Section 5 and Appendix A. The architecture of GlobalForge and its training objective are detailed in Section 4, and full hyperparameters, training schedules, and backbone settings are reported in Appendix D. The evaluation protocol, the baseline detectors, and the rationale for using official versus retrained weights are given in Section 6 and Appendix C, and additional results on standard benchmarks are provided in Appendix B.

References

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024. doi: 10.1109/OJSP.2023.3337714.

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

You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*, 2023.

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

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*, 2025.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13455–13465, 2025.

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

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 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023b.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Niesner, and Luisa Verdoliva. Raising the Bar of AI-generated Image Detection with CLIP . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4356–4366, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. doi: 10.1109/CVPRW63382.2024.00439. URL https://doi.ieeecomputersociety.org/10.1109/CVPRW63382.2024.00439.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18685–18694, 2025.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution ai-generated image detection by spectral learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18706–18717, 2025.

Jianpeng Ke and Lina Wang. Df-udetector: An effective method towards robust deepfake detection via feature restoration. *Neural Networks*, 160:216–226, 2023.

Despina Konstantinidou, Christos Koutlis, and Symeon Papadopoulos. Texturecrop: Enhancing synthetic image detection through texture-based cropping. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 1459–1468, 2025.

Chunxiao Li, Xiaoxiao Wang, Meiling Li, Boming Miao, Peng Sun, Yunjian Zhang, Xiangyang Ji, and Yao Zhu. Bridging the gap between ideal and real-world evaluation: Benchmarking ai-generated image detection in challenging scenarios. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20379–20389, 2025a.

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, 2025b.

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*, 2025c.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? *arXiv preprint arXiv:2505.12335*, 2025d.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10770–10780, 2024.

Ruiqi Liu, Yi Han, Zhengbo Zhang, Liwei Yao, Zhiyuan Yan, Jialiang Shen, ZhiJin Chen, Boyi Sun, Lubin Weng, Jing Dong, et al. Beyond artifacts: Real-centric envelope modeling for reliable ai-generated image detection. *arXiv preprint arXiv:2512.20937*, 2025.

Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, 54(1):1–41, 2021.

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.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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 Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

Anisha Pal, Julia Kruk, Mansi Phute, Manognya Bhattaram, Diyi Yang, Duen Horng Chau, and Judy Hoffman. Semi-truths: A large-scale dataset of ai-augmented images for evaluating robustness of ai-generated image detectors. *Advances in Neural Information Processing Systems*, 37:118025–118051, 2024.

Ziheng Qin, Yuheng Ji, Renshuai Tao, Yuxuan Tian, Yuyang Liu, Yipu Wang, and Xiaolong Zheng. Scaling up ai-generated image detection via generator-aware prototypes. *arXiv preprint arXiv:2512.12982*, 2025.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. *arXiv preprint arXiv:2410.11835*, 2024.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

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

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025.

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information fusion*, 64: 131–148, 2020.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020.

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

Haiwei Wu, Jiantao Zhou, and Shile Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *IEEE Transactions on Artificial Intelligence*, 2025.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024a.

Zhiyuan Yan, Jiangming Wang, Zhendong Wang, Peng Jin, Ke-Yue Zhang, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2(6):7, 2024b.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.

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

A REALDEG-BENCH: CONSTRUCTION DETAILS

Section 5 of the main text describes the design goal and degradation operator pool of RealDeg-Bench. This appendix provides additional details on the construction pipeline, including source data selection, strength sets of each operator, single-operator and compound sampling protocols, and dataset statistics.

Source data. RealDeg-Bench is constructed from a T2I-CoReBench (Li et al., 2025c)-derived image pool. For fake images, we randomly sample 3,978 images with a balanced distribution across generator-specific subsets. For real images, we use 3,375 images from the corresponding real-image split.

Degradation operators and strength sets. The seven atomic operators correspond to common distortions in real propagation chains. JPEG compression models platform re-encoding; resizing models resolution resampling; Gaussian blur models defocus and detail loss after repeated compression or reposting; Gaussian noise models capture and transmission noise; and brightness, contrast, and saturation simulate appearance shifts caused by platform-side rendering, screenshot re-sharing, or post-processing. Each operator is associated with a discrete set of strength levels covering a range from mild to severe:

- JPEG quality factor: $Q \in \{90, 80, 70, 60, 40\}$
- Gaussian blur standard deviation: $\sigma \in \{0.5, 1, 2, 3, 5\}$
- Resize scale: $s \in \{0.9, 0.7, 0.5, 0.3, 0.2\}$
- Gaussian noise variance: $v \in \{0.0005, 0.001, 0.002, 0.005, 0.01\}$
- Brightness shift: $b \in \{-0.2, -0.1, 0.1, 0.2\}$
- Contrast shift: $c \in \{-0.3, -0.2, 0.1, 0.2\}$
- Saturation factor: $f \in \{0.6, 0.8, 1.3, 1.5\}$

Single-operator protocol. For each of the seven single-operator conditions, every image is transformed by one specified operator. For each image, the strength of that operator is independently and uniformly sampled from its discrete strength set, producing one degraded version per image. Different images within the same condition therefore receive different randomly drawn strengths.

Multi-step compound chain sampling. For compound degradation, we construct an N-step chain for $N \in \{1, 2, 3, 4, 5\}$. For each image, the chain is formed by sampling N operators from the seven-operator pool and applying them sequentially. This design reflects realistic propagation, where an image may undergo multiple rounds of compression, resizing, color adjustment, or other platform-side processing as it traverses successive online channels. For every operator instance in the chain, its strength is independently and uniformly drawn from the corresponding strength set, so different occurrences of the same operator may apply different strengths.

Dataset statistics. RealDeg-Bench contains 13 conditions in total: clean, seven single-operator conditions, and five compound conditions ($N = 1, \ldots, 5$). Each condition contains 3,375 real images and 3,978 fake images, yielding 7,353 images per condition and 95,589 images in total. The sampled operator sequence and the strength drawn for every operator instance are recorded and stored together with the degraded images, so all evaluations on RealDeg-Bench are deterministic and reproducible.

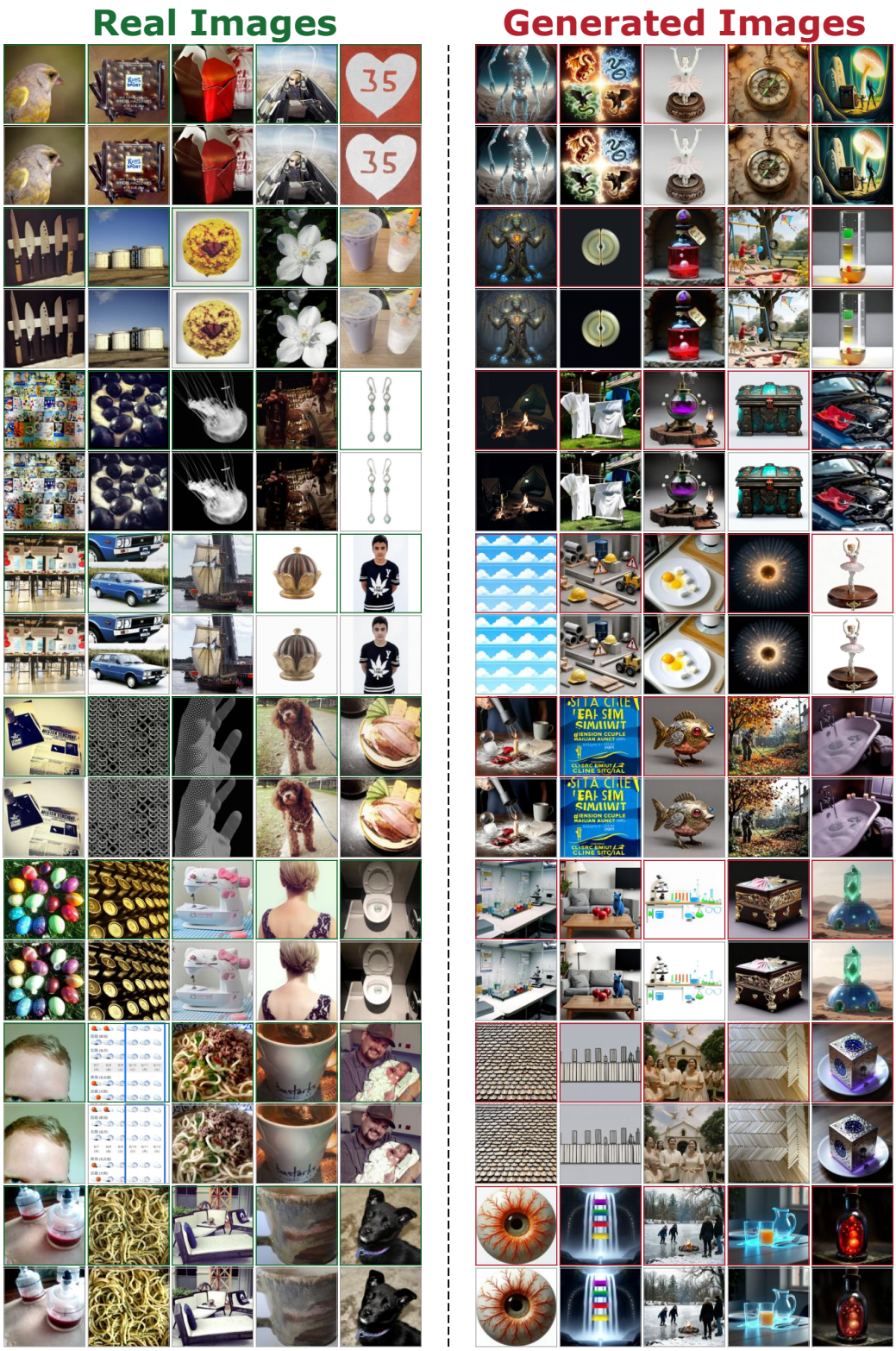

Figure 7: **Example samples from *RealDeg-Bench*.** Rows alternate between clean images from our dataset (odd rows) and their corresponding degraded versions (even rows), illustrating the diversity of generators, content, and degradation types covered by the benchmark.

Table 5: Standard benchmark results. We report Balanced Accuracy (B.Acc, %) on clean images and Degraded Accuracy (D.Acc, %) under one randomly sampled degradation operator from RealDeg-Bench. Bold indicates the best Avg; underline indicates the second-best Avg.

Method	AIGCDetect		GenImage		UnivFD		DRCT		Synthbuster		EvalGEN		Avg	
	B.Acc	D.Acc	B.Acc	D.Acc	B.Acc	D.Acc	B.Acc	D.Acc	B.Acc	D.Acc	B.Acc	D.Acc	B.Acc	D.Acc
GAPL (Qin et al., 2025)	97.43	87.82	96.57	86.84	90.94	83.89	96.64	86.85	95.75	87.09	97.53	89.36	**95.81**	<u>86.97</u>
B-Free (Guillaro et al., 2025)	84.66	79.93	89.63	82.80	87.77	81.57	99.25	95.08	95.74	92.23	94.60	89.83	91.94	86.91
DDA (Chen et al., 2025)	81.52	77.48	88.91	81.96	69.38	71.56	97.03	88.95	95.38	88.24	95.41	85.74	87.94	82.32
Effort (Yan et al., 2024b)	71.22	65.42	86.29	74.87	55.07	58.06	89.09	74.87	88.93	76.09	97.08	78.12	81.28	71.24
DRCT (Chen et al., 2024)	67.06	60.74	78.75	68.71	66.77	62.62	97.01	79.69	80.99	72.82	81.91	78.60	78.75	70.53
AIDE (Yan et al., 2024a)	84.00	70.01	88.56	76.75	77.85	64.39	59.24	63.09	75.40	70.88	59.54	70.28	74.10	69.23
SAFE (Li et al., 2025b)	48.36	63.62	47.74	74.37	49.69	57.43	50.31	68.78	51.56	75.65	50.15	69.18	49.64	68.17
UnivFD (Ojha et al., 2023)	56.46	69.81	62.51	64.20	48.28	71.89	69.50	61.03	65.74	65.61	73.78	60.14	62.71	65.45
FatFormer (Liu et al., 2024)	82.13	62.05	71.50	63.86	86.71	60.89	53.88	62.13	69.08	70.18	56.59	66.07	69.98	64.20
C2P-CLIP (Tan et al., 2025)	79.95	68.33	71.40	60.42	87.30	74.38	54.58	56.17	69.42	61.24	57.72	63.53	70.06	64.01
Aligned (Rajan et al., 2024)	55.70	53.83	57.50	54.93	60.48	58.07	54.87	56.82	54.54	54.24	65.76	58.35	58.14	56.04
NPR (Tan et al., 2024)	64.79	58.68	68.49	60.94	51.99	51.89	69.15	60.45	73.74	64.18	68.23	58.70	66.07	59.14
GlobalForge	93.30	87.96	96.76	90.64	86.90	85.13	99.22	93.27	99.41	95.58	97.14	89.04	<u>95.45</u>	**90.27**

B Supplementary Results on Standard Benchmarks

In addition to the real-degradation and in-the-wild evaluations in the main text, we also report full numbers on common standard benchmarks as a reference for conventional detection capability. Table 5 summarizes B.Acc and D.Acc on six public standard benchmarks: AIGCDetect (Zhong et al., 2023), GenImage (Zhu et al., 2023), UnivFD (Ojha et al., 2023), DRCT (Chen et al., 2024), Synthbuster (Bammey, 2024), and EvalGEN (Chen et al., 2025). B.Acc is measured on clean images; D.Acc is measured after applying one randomly sampled degradation operator from our RealDeg-Bench to each benchmark. These results show that our robustness gains do not come at the cost of standard benchmark performance: GlobalForge achieves the best average D.Acc of **90.27%** and the second-best average B.Acc of **95.45%** across all methods.

Table 6: Comparison of **Official Weights** and **Our Retrained Weights** on in-the-wild benchmarks. We report Balanced Accuracy (B.Acc, %) on each dataset group and the equal-weighted average over the 8 groups. **Bold** marks the better weights in Average; all `Ours` rows are trained from scratch on the same MSCOCO (Lin et al., 2014) + AIGI split as our model.

Method	Weights	Chameleon	SynthWildx	WildRF	AIGIBench	CO-SPY	RR-Dataset	BFree-Online	Real-Chain	Average
AIDE (Yan et al., 2024a)	Official	59.35	57.65	62.49	58.69	22.23	52.47	51.76	40.82	50.68
	Ours	64.48	68.89	68.91	64.66	54.17	56.91	67.35	45.09	**61.31**
SAFE (Li et al., 2025b)	Official	52.50	49.93	51.26	50.40	0.83	49.81	50.00	36.45	42.65
	Ours	51.80	53.40	52.37	53.67	98.05	56.76	55.19	63.09	**60.54**
NPR (Tan et al., 2024)	Official	61.26	54.58	64.22	63.25	78.79	56.99	66.06	49.55	**61.84**
	Ours	62.73	60.91	60.36	62.22	71.26	57.24	55.95	45.55	59.53
Effort (Yan et al., 2024b)	Official	52.41	49.86	63.13	57.40	47.07	49.59	53.49	46.18	52.39
	Ours	69.03	55.80	48.21	62.30	93.65	52.61	59.30	65.36	**63.28**
FatFormer (Liu et al., 2024)	Official	50.88	49.82	49.95	50.23	96.14	50.58	49.85	50.57	56.00
	Ours	62.10	57.92	61.04	58.42	79.39	50.44	54.16	53.45	**59.62**
UnivFD (Ojha et al., 2023)	Official	50.18	51.12	54.98	52.86	9.10	46.21	48.81	35.82	43.64
	Ours	74.46	55.29	50.00	66.17	72.48	49.41	69.68	62.27	**62.47**

C Comparison between Official Weights and Our Retrained Weights

In the in-the-wild evaluation of Section 6.3, the 12 baselines are evaluated under two settings depending on whether their training pipelines can be cleanly transferred to a unified protocol.

For C2P-CLIP (Tan et al., 2025), DRCT (Chen et al., 2024), Aligned (Rajan et al., 2024), B-Free (Guillaro et al., 2025), DDA (Chen et al., 2025), and GAPL (Qin et al., 2025), we follow standard community practice and use their officially released weights. DRCT, Aligned, B-Free, and DDA are tightly coupled with their own proposed alignment datasets, and the data curation and alignment procedure is an integral part of the training pipeline that cannot be cleanly decoupled from the detector itself. C2P-CLIP and GAPL adopt specialized training paradigms (image text contrastive learning with category common prompts and prototype mapping over canonical generators,

Table 7: Complete training setup of the two model variants. Fields marked with † were absent from the original table and are derived directly from the source code.

Field	ours-d2	ours-d3
Backbone & LoRA adaptation		
Backbone (frozen)	DINOv2-L (`dinov2-large`)	DINOv3-L (`dinov3-large`)
Input / patch size†	224×224, patch 14	224×224, patch 16
LoRA†	r=16, α=32, dropout 0, on Q/K/V of all 24 layers	
Optimisation		
Optimizer†	AdamW (β_1=0.9, β_2=0.999, ϵ=10^{-8}), weight decay 0.01	
LR (peak / min)	4×10^{-4} / 2×10^{-5}	5×10^{-5} / 1×10^{-6}
LR schedule†	Half-cycle cosine decay, 1 warm-up epoch, fp32, no grad clipping	
Schedule, batch & module hyperparameters		
GPUs	1×A100-80 GB	4×A100-80 GB
Epochs (scheduled / logged)	10 / 8	8 / 6
Batch / GPU, accum, effective	128, 16, 2048	128, 4, 2048
Samples per epoch†	80 000 (40k real + 40k fake, capped from 118,287/class); seed 0, single run	
LIB†	k=3, σ=1.0; gate β as a single learnable scalar	
GSR†	w_{gsr}=3 (Chebyshev distance threshold)	
DCS loss†	λ_{dcs}=0.01, τ=0.07	
DCS augmentation†	JPEG $Q\in(20, 80)$; blur $\sigma\in(0.5, 1.5)$ with 7×7 kernel; color jitter 0.4 / hue 0.06	

Software / hardware. All experiments run on NVIDIA A100 80 GB PCIe, Python 3.10.18, PyTorch 2.8.0 + cu126 (CUDA 12.6).

respectively) that are incompatible with a unified binary classification pipeline. Retraining any of them on a different data source or under a different objective would fundamentally alter the approach being evaluated, so we adopt the released checkpoints to faithfully represent each method's original design.

The remaining six baselines, NPR (Tan et al., 2024), UnivFD (Ojha et al., 2023), FatFormer (Liu et al., 2024), SAFE (Li et al., 2025b), AIDE (Yan et al., 2024a), and Effort (Yan et al., 2024b), are retrained from scratch following the experimental settings of DDA (Chen et al., 2025), sharing the same training data, splits, and evaluation pipeline as GlobalForge. Aside from minimal modifications required to plug into the unified pipeline, we do not alter the core code of any baseline.

Across both the official-weights and retrained-weights settings, we retain each baseline's original backbone rather than transplanting every method onto a unified backbone, because most of these detectors are designed around backbone-specific properties, and forcing them onto a uniform feature extractor would invalidate those design choices and no longer reflect the method as proposed.

As shown in Table 6, our retrained weights outperform the officially released weights on the average of the 8 in-the-wild benchmark groups for most methods, confirming that the gains come from better data quality and training uniformity rather than any unfair advantage.

D Implementation Details

For reproducibility, this section consolidates the complete training and inference setup of GlobalForge. GlobalForge follows a "**pretrained vision backbone + lightweight structural modules + joint training objectives**" paradigm. The input image is patchified and fed into the pretrained Transformer encoder; LIB and GSR are inserted sequentially after the backbone output; and the classification head produces the real/fake decision. Unlike many methods that rely on additional restoration branches, frequency-domain branches, or test-time ensembling, GlobalForge's main inference path remains a single forward pass, and the additional modules all operate in feature space. Hence, the key training details lie not in a complex data flow, but in the stable joint optimization of local suppression, global reasoning, and the degradation-invariance constraint.

Compute resources. All experiments were run on NVIDIA A100 80GB PCIe GPUs. The main GlobalForge variants used 1 A100 GPU for ours-d2 and 4 A100 GPUs for ours-d3. Training ours-d2 took approximately 9 GPU-hours, and training ours-d3 took approximately 24 GPU-hours.

Optimizer groups. All trainable parameters are managed by a single AdamW optimizer and partitioned into two parameter groups by weight-decay policy: a *decay* group (projection matrices and linear layers) with weight decay 0.01, and a *no-decay* group (biases, layer-norm scales, and the LIB gate scalar β) with weight decay 0.

Construction of degraded views. The degradation-aware contrastive structural loss $\mathcal{L}_{\text{DCS}}$ requires an online-constructed degraded view for each training sample. Following Section 4.3 in the main text, this view is constructed by random JPEG compression, random Gaussian blur, and random color perturbation. The goal is to create sufficiently diverse appearance perturbations for the same semantic content during training, so that the model is forced to learn representations that are stable across degradations. The image representation $\mathbf{z}_i$ is obtained by average-pooling the patch-token outputs from the final backbone layer.

Alternative-analysis variants. To support the alternative analysis in Table 4, we train four additional variants following the same implementation pipeline as the main model, with only the tested mechanism changed. *Aug* removes LIB, GSR, and $\mathcal{L}_{\text{DCS}}$, and trains the classifier with standard random degradation augmentation only. *Blur* replaces the feature-domain LIB with input Gaussian smoothing, applying a 3×3 Gaussian blur with $\sigma = 1.0$ to the image input during both training and evaluation. *FFT* replaces LIB with input frequency low-pass filtering, using a cutoff ratio of 0.25 during both training and evaluation. *Global Attention* keeps the attention module global but disables the GSR, thereby testing an unconstrained global receptive-field alternative. All variants are evaluated on the one-step compound split of RealDeg-Bench, denoted as RealDeg-1Step in Table 4.

Table 7 records the key hyperparameters actually used by GlobalForge in the main experiments, which directly reproduce the experimental setup and are useful for comparisons with other methods.

Dataset and asset licenses. All datasets and pretrained models are used in compliance with their respective licenses and solely for non-commercial academic research. DINOv2 (Oquab et al., 2024) is used under Apache 2.0; DINOv3 (Siméoni et al., 2025) under Meta's non-commercial research license; GenImage (Zhu et al., 2023) and Synthbuster (Bammey, 2024) under CC BY-NC-SA 4.0; DRCT (Chen et al., 2024), EvalGEN (Chen et al., 2025), and T2I-CoReBench (Li et al., 2025c) under Apache 2.0; LoRA (Hu et al., 2022) under MIT. AIGCDetect (Zhong et al., 2023) and UnivFD (Ojha et al., 2023) do not publish explicit licenses and are accessed through their official repositories for research use only.

E Supplementary Visual Examples

Beyond Figure 3 in the main text, we provide two additional qualitative views: (i) typical success and failure cases of our detector, and (ii) side-by-side attention heatmaps on clean and degraded inputs for several detectors. Both views are interpreted under the same lens used throughout the paper: GlobalForge is designed to base its decision on long-range structural evidence rather than on any local artifact signature, and these visualizations are meant to make that design choice empirically visible.

Qualitative analysis of success and failure cases. Fig. 8 shows representative success and failure cases together with the model's Grad-CAM attention. The highlighted regions mark tokens whose contribution to the final score is large; under the GSR design (Section 4.2), each such token has already aggregated information from distant patches outside its own local window, so the highlights should be read as anchors of a long-range structural judgement rather than as isolated local cues. In the successful cases (top row), these anchors are spread across multiple regions of the image instead of clustering on a single local patch, and the predicted probabilities are confident on both sides ($p \leq 0.12$ for real and $p \geq 0.90$ for fake). The failure cases (bottom row) sit close to the decision boundary and share a common pattern: the image offers little exploitable long-range structure. Real images with smooth or low-texture content (e.g. flat color regions, close-ups of repetitive patterns) provide too

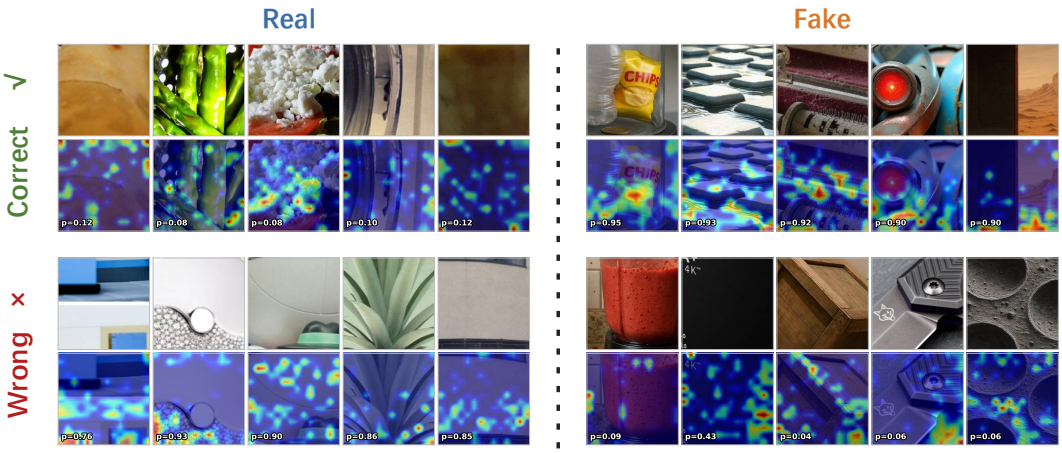

Figure 8: **Qualitative success and failure cases of GlobalForge.** The figure uses a 2×2 block layout (correct vs. wrong × real vs. fake). Each block contains five columns; the input image is on top and the Grad-CAM heatmap (annotated with the predicted fake probability p) is on the bottom. The four blocks are: correctly classified real images (top-left, $p \leq 0.12$), correctly classified fake images (top-right, $p \geq 0.90$), real images misclassified as fake (bottom-left, $p > 0.5$), and fake images misclassified as real (bottom-right, low-to-borderline p).

few cross-region cues, so a small number of accidental responses is enough to push the score over the boundary and the image ends up labelled as fake. Fake images dominated by natural-looking objects or strong photographic priors (e.g. realistic lighting, plausible material reflections) instead present a globally coherent appearance whose long-range structure does not differ visibly from a real photograph, and the detector misses them. The remaining errors therefore concentrate on samples where the global-structure signal is weak or already close to that of natural images, which is consistent with our design principle of relying on global structural coherence rather than on any specific local artifact signature.

Stability of spatial focus under degradation. Figure 9 compares the spatial responses of GlobalForge with six representative baselines on four image pairs, where each pair is a clean image and a degraded counterpart obtained from a single perturbation. Within a pair, the two heatmaps share a normalization range, so visual differences read as direct evidence about the stability of the method's focus. Across all four pairs, the baselines change a lot after degradation: their attended regions drift to unrelated areas, become diffuse, or collapse onto a few isolated points that no longer correspond to the regions used on the clean image. This is the visual signature of detectors whose decision is anchored on local artifacts: once a perturbation alters the local statistics, the spatial focus has nothing stable to hold on to. Our method instead keeps its attention on essentially the same regions before and after degradation, with only minor changes in intensity. We attribute this to the GSR design (Section 4.2), which forces every token to aggregate evidence from distant patches outside its local window: the resulting decision is supported by a long-range structural pattern that survives a single local perturbation, and the contrastive structural loss further pulls clean and degraded representations together so that the global structural cues stay aligned. This matches the quantitative results in Section 6.2: as the input statistics shift, baseline detectors lose their local discriminative cues, while GlobalForge still anchors its predictions on the global structural pattern, which is the underlying reason it degrades more gracefully.

F Broader Impact

The development of robust AI-generated image (AIGI) detection methods has important societal, ethical, and technical implications. Our work contributes to mitigating the risks of synthetic visual content, including misinformation, impersonation, and non-consensual deepfakes, by enabling more reliable verification of image authenticity under realistic online propagation. Images shared on social platforms are often compressed, resized, blurred, and re-shared before inspection; by explicitly targeting these degraded in-the-wild conditions, GlobalForge provides a practical framework for detecting AI-generated images beyond clean benchmark settings. This can strengthen trust in digital

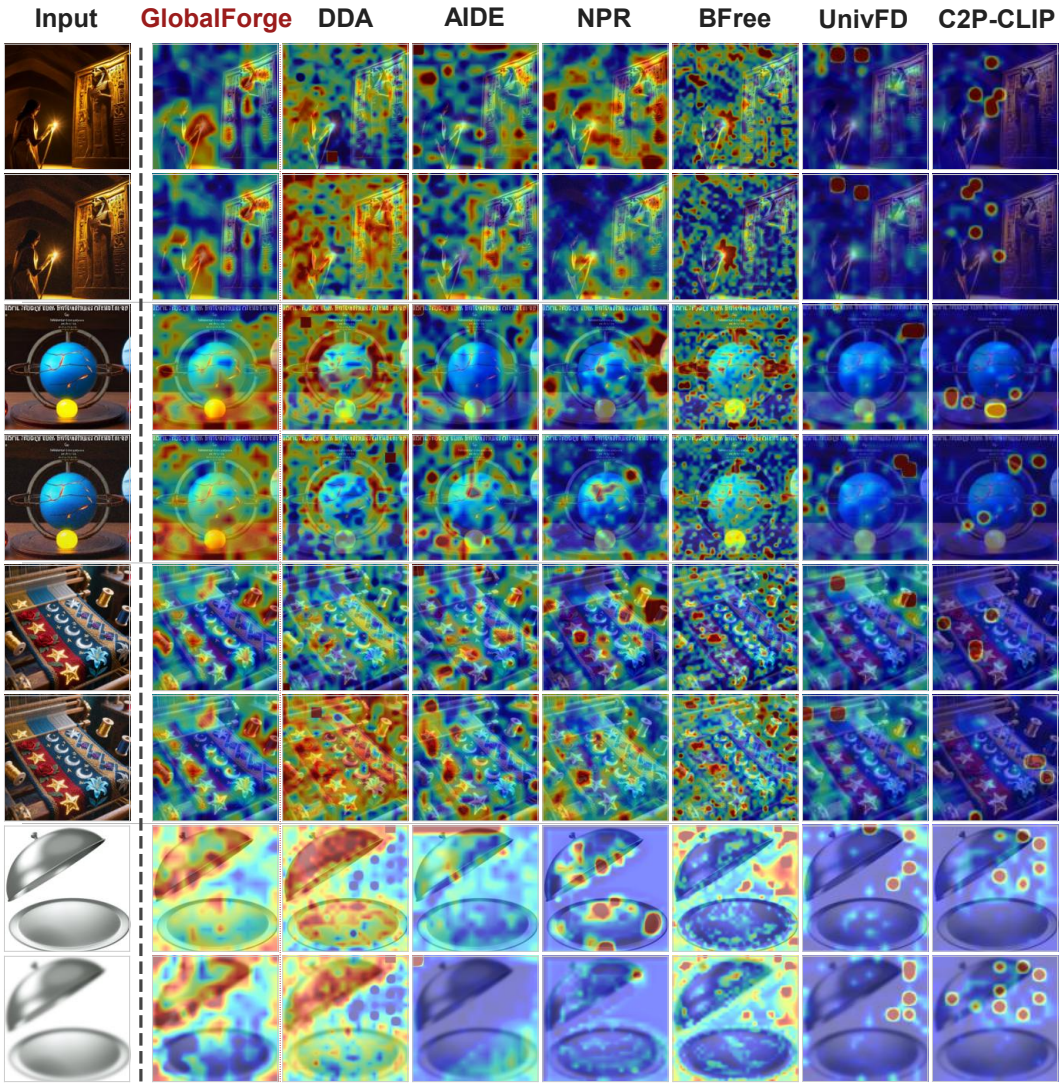

Figure 9: **Attention stability under degradation: GlobalForge versus six representative baselines.** Each pair of consecutive rows is one image, shown clean (top) and degraded (bottom). Columns from left to right are: the input image, GlobalForge (ours), and six competing detectors (DDA, AIDE, NPR, B-Free, UnivFD, C2P-CLIP). For each method, the two heatmaps in a pair use the same normalization range, so a visible change in the highlighted regions between the two rows is a direct sign that the method's spatial focus is unstable.

media, support content moderation efforts, and assist forensic or journalistic investigations involving visual evidence.

This research also aligns with broader efforts to establish trustworthy multimedia ecosystems. By shifting detection from fragile local generator artifacts toward degradation-stable global structural evidence, GlobalForge advances a robustness-oriented detection mechanism that is better matched to real-world deployment conditions. The proposed RealDeg-Bench further supports this goal by enabling fine-grained evaluation of detector robustness under common propagation operations and compound degradation chains. We hope these contributions can facilitate more reliable auditing of AI-generated visual content while promoting public awareness of synthetic media risks.

At the same time, AIGI detection should be used with care. No detector is fully reliable, and incorrect predictions may lead to false accusations, overlooked synthetic content, or misplaced confidence in automated decisions. Therefore, we recommend treating detector outputs as supporting evidence rather than definitive judgments, especially in high-stakes settings such as journalism, legal analysis,

or content moderation. Human review, provenance metadata, and contextual information should be considered alongside model predictions.

We encourage interdisciplinary collaboration among researchers, platform operators, journalists, forensic analysts, ethicists, and policymakers to ensure that AIGI detection technologies serve as safeguards for digital media authenticity and public trust.

G Limitations and Future Work

While GlobalForge demonstrates strong robustness across controlled degradation settings and in-the-wild AIGI detection benchmarks, several limitations and opportunities for future work remain.

Limitations. First, although RealDeg-Bench covers a broad set of common propagation degradations, real-world image circulation is more diverse and continuously evolving. Platform-side compression policies, editing tools, screenshot behaviors, and re-sharing pipelines may change over time, which can introduce degradation patterns beyond the current benchmark. Second, GlobalForge is designed around the assumption that global structural cues are more stable than local artifacts. This assumption is well supported by our experiments, but future generative models may reduce such structural inconsistencies as synthesis quality improves. Third, while LIB and GSR use lightweight designs, their key hyperparameters, such as the local suppression strength and the masking window, are still manually specified. More adaptive designs may further improve robustness across datasets and resolutions.

Future Directions. Future work could extend RealDeg-Bench with dynamic re-calibration for newly emerging platform pipelines and more diverse user-side editing operations. Another promising direction is to study whether global structural cues remain reliable for more advanced generators, localized image editing, and multimodal generation scenarios. In addition, adaptive local suppression and structure-aware reasoning modules could be explored to reduce manual design choices while preserving the robustness benefits of GlobalForge. Finally, extending the framework from image-level AIGI detection to video-generated content detection may further test whether degradation-stable structural reasoning can generalize beyond static images.