BEYOND NATURAL IMAGES: RETHINKING AI-GENERATED IMAGE DETECTION IN DOCUMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image detection has attracted increasing attention, but existing evaluations mainly focus on natural images, leaving AI-generated document images largely underexplored. This omission is concerning because documents often appear in sensitive real-world scenarios, such as invoices, expense reports, certificates, and medical records. In this paper, we first construct a controlled diagnostic benchmark, **AIGDoc-Pilot**, and reveal that existing detectors suffer substantial performance degradation on AI-generated document images, with the mean AUC dropping by more than 7%. Based on this, we further reveal two document-specific properties behind this gap: generation artifacts exhibit strong spatial inconsistency across local regions, and text density significantly affects real-synthetic separability, where text-dense regions offer stronger discriminative evidence. Motivated by these findings, we construct **AIGDoc**, a larger document-centric dataset containing diverse real-world documents and AI-generated counterparts produced by multiple advanced generation and editing models. Extensive experiments on AIGDoc demonstrate that existing detectors still struggle to reliably identify AI-generated documents, while document-based training partially narrows the gap. Together, these results offer valuable insights for developing dependable and generalizable detectors in document-centric scenarios. **The code and datasets will be made publicly available upon acceptance of the paper.**

1 INTRODUCTION

With the continuous advancement of AI-generated image technologies Ye et al. (2023); Wu et al. (2024), realistic images can now be produced with greater ease and efficiency, thereby enriching people's daily lives. However, this development has also made such images increasingly difficult to distinguish from authentic ones, posing significant potential risks to society, the economy, and politics Wang et al. (2024; 2025a; 2026a); Hu et al. (2023; 2026). Many researchers have begun to focus on training highly generalizable detectors for AI-generated images, including methods based on artifact representations Li et al. (2025a); Lin et al. (2025); Yan et al. (2026) and pre-trained models Zhou et al. (2025a); Liu et al. (2025a), which have been applied effectively to real-world scenarios. Despite their promising performance in real-world scenarios, existing studies predominantly focus on natural images, while risks posed by AI-generated document images remain largely overlooked.

Traditionally, document forgery has often relied on manual editing tools such as Photoshop. Many studies have explored this problem within the framework of document tampering detection Luo et al. (2025); Qu et al. (2026b;c). However, recent generative models can synthesize entire document images or modify existing ones with high visual realism, introducing forgery patterns that may not be adequately captured by conventional tampering detection methods. These cases blur the boundary between document tampering detection and AI-generated image detection, and naturally raise a critical question: *Can existing detectors really cover AI-generated document images?*

To comprehensively investigate this problem, we first construct a pilot benchmark, termed **AIGDoc-Pilot**. It contains real-world camera-captured [1] document images from multiple practical scenarios,

[1] Existing detectors are typically trained on camera-captured images. Accordingly, digital images, rendered images, and similar data are regarded as cross-type images. In this study, we restrict the dataset to camera-captured images to ensure source consistency.

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

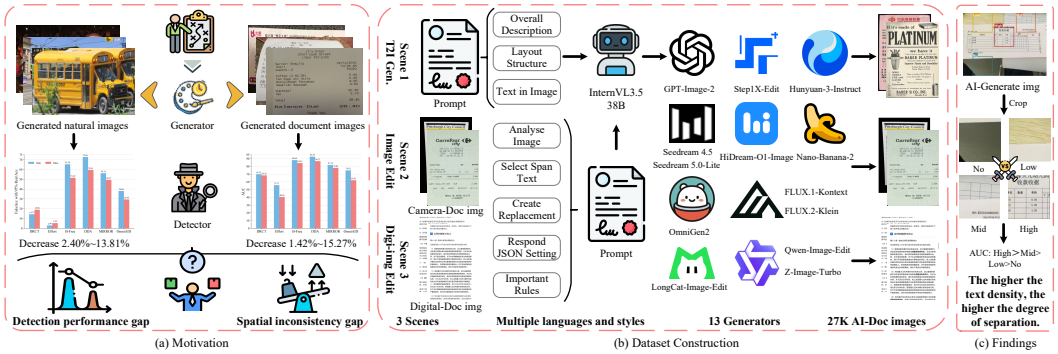


Figure 1: **(a) Diagnosis and Findings.** Using the controlled diagnostic benchmark AIGDoc-Pilot, we reveal that existing detectors suffer substantial performance degradation on AI-generated document images. Further diagnosis shows that generated document artifacts exhibit strong spatial inconsistency across local regions, while text-dense regions provide stronger discriminative evidence for real-synthetic separability. **(b) Datasets and Evaluations.** We construct AIGDoc, a large-scale benchmark for AI-generated document image detection, comprising 27K synthetic images across 17 subsets, covering 13 generation methods from three categories.

together with their AI-generated counterparts produced by image-to-image generation under different noise levels. For comparison, we also collect an equal number of camera-captured natural images and generate corresponding synthetic images using the same pipeline. Experimental results demonstrate that existing detection methods encounter significant challenges when evaluated on AIGDoc-Pilot. Key insights are summarized as follows: **i) A pronounced gap between natural and document images.** Under identical generation conditions, the same detectors suffer a substantial accuracy drop on AI-generated document images compared with AI-generated natural images. **ii) Strong spatial inconsistency of artifacts in AI-generated document images.** Compared with natural images, AI-generated document images demonstrate significantly weaker spatial coherence of generation artifacts. Detectable forgery evidence is often concentrated in limited local regions, whereas others contain only subtle artifact traces, resulting in substantial discrepancies in detection performance across distinct image patches. **iii) Text density enhances authentic-synthetic separability.** We observe that regions with denser textual content enable detectors to better distinguish AI-generated document images from real ones, suggesting that discriminative artifact cues in document images are closely associated with high-frequency character boundaries and layout structures.

Meanwhile, to facilitate systematic evaluation and alleviate the observed limitations of existing AI-generated image detectors on document images, we propose the **AIGDoc** dataset, which includes diverse real-world document images and AI-generated samples produced by multiple advanced generation and editing models. We show that incorporating document images during training can partially reduce the natural-document gap, underlining the need for document-centric datasets, evaluation protocols, and detection methods. Overall evaluation and datasets are shown in Figure 1.

The main contributions of this paper are as follows:

- To the best of our knowledge, this is the first systematic investigation of AI-generated image detection in document-centric scenarios. Through AIGDoc-Pilot, we reveal a pronounced performance gap between natural images and document images, highlighting an overlooked limitation of existing AI-generated image detectors.
- Our findings show that generation artifacts in document images exhibit strong spatial inconsistency, and text-dense regions provide stronger discriminative artifact cues for distinguishing real and AI-generated documents. The simple text-density-guided preprocessing during training significantly boosted model performance, surpassing the state-of-the-art method by 8.9% in mAcc.
- We construct AIGDoc, a large-scale AI-generated document image dataset comprising 27k AI-generated images across 17 subsets, covering 13 generation methods from three categories. AIGDoc provides a practical benchmark for systematic evaluation, helping bridge the gap between AI-generated image detectors and document-centric detection scenarios.

2 Related Work

2.1 AI-Generation Image Detection

The rapid advancement of AI generative models has introduced serious risks to visual authenticity. To address these concerns, recent studies have focused on developing AI-generated image detectors with strong generalization ability. Artifact-based approaches typically exploit generation traces such as frequency-domain anomalies Luo et al. (2021); Tan et al. (2024a); Yan et al. (2026), upsampling artifacts Tan et al. (2024b), local patch inconsistencies Chen & Yang (2021); Cavia et al. (2024), and reconstruction residuals Wang et al. (2023); Chen et al. (2024); Chu et al. (2025). In recent years, leveraging the transferable and robust representations of pre-trained models has become an important direction for improving detection performance, which includes feature-based methods Ojha et al. (2023); Koutlis & Papadopoulos (2024); Zhang et al. (2025a); Zhou et al. (2025b) and fine-tuning-based methods Liu et al. (2024); Tan et al. (2025); Yan et al. (2025b;a); Liu et al. (2026). For example, VIB-Net Zhang et al. (2025a) applies a variational information bottleneck to suppress task-irrelevant information, NS-Net Yan et al. (2025b) constructs a semantic NULL space to remove semantic interference from visual features, and DGS-Net Yan et al. (2025a) introduces distillation-guided gradient surgery to suppress harmful gradients while preserving transferable pre-trained representations. In addition, several studies improve generalization from the data perspective, such as data augmentation Li et al. (2025b) and the construction of less biased training datasets Chen et al. (2025); Guillaro et al. (2025). Despite these efforts and the promising generalization achieved on natural images, AI-generated document image detection remains largely unexplored.

2.2 AI-Counterfeited Document Image Detection

With the maturation of image-editing tools such as Photoshop, many researchers have devoted considerable effort to document tampering detection, in which high-frequency inconsistencies around tampering boundaries are often exploited for both detection and localization Luo et al. (2025); Song et al. (2025); Qu et al. (2026b;a;c), achieving promising performance. However, the rise of AI-generated content Wu et al. (2025a;b); Black Forest Labs et al. (2025) has greatly expanded editing capabilities beyond conventional manual manipulation. In particular, image editing can remove the sharp boundary artifacts and local high-frequency inconsistencies that traditional tampering detectors rely on, posing a challenge to existing document forgery detection methods.

Notably, existing AI-generated image detection research has rarely considered document images, leaving a substantial gap in the evaluation of document-centric counterfeits. To address this gap, we conduct the controlled diagnostic study on **AIGDoc-Pilot** and introduce the larger real-world benchmark **AIGDoc**. The most closely related work is TextFake Zhang et al. (2026), which reports detector failures on AI-generated text-rich images but mixes diverse image types and generation settings, making it difficult to determine whether the observed degradation comes from the document domain itself. In contrast, our work constructs a controlled comparison under the same generation pipeline, thereby enabling a direct measurement and analysis of the document-domain gap. We hope that our study can help bridge the gap between natural-image-oriented and document-centric forgery detection, providing a foundation for future research on reliable AI-generated document image detection.

3 Diagnostic Study on AI-Generated Document Images

3.1 AIGDoc-Pilot Datasets

Real Image Collection. We observe that the training data used by existing AI-generated image detectors mainly consist of camera-captured natural images, such as MSCOCO Lin et al. (2014) and ImageNet Deng et al. (2009). In contrast, document images are more diverse in acquisition forms, including camera-captured images, screenshots, and born-digital electronic documents. Directly comparing such heterogeneous document images with camera-captured natural images may introduce confounding cross-domain factors and lead to inaccurate performance evaluation. Therefore, we collect candidate document images from SRD Auad et al. (2024), M6DocCheng et al. (2023), and the ICDAR competition dataset Yu et al. (2023). We then manually filter out images that are

Table 1: **Detection Performance Gap between AI-Generated Natural and Document Images on AIGDoc-Pilot.** For each detector, the decision threshold is calibrated to maintain 95% Real Acc. Gray rows denote results on document images, while white rows denote results on natural images.

Method	SDv2.1-VAE	HiDream-E1-Full	FLUX.1-dev				Qwen-Image		SDv3.5-Large		Mean	
			0.2	0.4	0.6	0.8	0.2	0.4	0.2	0.4	F.Acc.	AUC
DRCT	13.21	15.41	6.61	9.17	17.98	29.54	11.56	17.06	10.64	12.48	14.37	**69.78**
	15.23	9.54	7.89	11.56	19.82	35.96	24.06	37.19	14.63	13.70	**18.96**$_{+4.59}$	68.36$_{-1.42}$
Effort	1.83	1.10	0.55	2.02	4.40	7.89	4.77	4.77	1.10	0.55	2.90	**55.73**
	1.65	1.10	1.10	2.20	5.87	15.96	10.94	12.50	0.75	0.94	**5.30**$_{+2.40}$	40.46$_{-15.27}$
B-Free	98.53	70.83	27.89	22.94	32.11	51.38	77.98	82.75	93.21	95.41	**65.30**	**88.08**
	91.38	49.36	17.06	19.63	35.60	41.83	62.81	70.62	59.85	66.79	51.49$_{-13.81}$	84.23$_{-3.85}$
DDA	96.88	67.52	40.55	49.36	62.39	63.30	84.40	84.59	88.07	89.54	**72.66**	**92.54**
	95.23	51.01	22.94	24.04	33.76	58.17	69.06	86.25	71.11	82.55	59.41$_{-13.25}$	86.75$_{-5.80}$
MIRROR	43.12	44.95	13.21	56.51	75.23	90.28	54.86	65.14	44.77	73.76	**56.18**	**81.78**
	50.09	40.73	12.84	24.04	44.04	55.41	50.31	93.12	39.96	83.11	49.37$_{-6.82}$	78.05$_{-3.73}$
OmniAID	48.44	55.96	10.28	22.57	30.83	62.94	29.72	35.96	27.89	55.41	**38.00**	**74.82**
	19.08	44.40	9.36	7.71	9.72	21.28	32.50	58.13	34.52	57.79	29.45$_{-8.55}$	62.03$_{-12.79}$

not captured by cameras, resulting in 545 camera-captured document images, which covers multiple document types such as books, receipts, and tickets. Meanwhile, for comparison, we randomly select the same number of natural images from Open Images V7 as the reference natural-image set. Under this constraint, both natural and document images share the same acquisition modality, where the primary difference lies in the visual content rather than the acquisition source.

AI-Generated Image Construction. To preserve the visual distribution of the original images, we construct AI-generated images using image-to-image generation. Specifically, we employ recent generative models, including HiDream-E1-Full Cai et al. (2025b), FLUX.1-dev Black Forest Labs, Qwen-Image Wu et al. (2025a), and SDv3.5-Large Esser et al. (2024). Within each subset, every generated image is paired with its corresponding source real image. To reduce bias caused by a single noise strength, we generate images under multiple noise levels. In addition, we also use the VAE of SDv2.1 Rombach et al. (2022), as it has been widely involved in the training data of existing detectors. To prevent shortcut bias caused by image format differences, all generated images are saved in JPEG format with a quality factor of 95 before evaluation. By restricting both domains to camera-captured images and applying identical generation and evaluation pipelines, we substantially reduce confounding effects from acquisition modality and generation settings, enabling us to assess the impact of domain shift on the model. More details can be found in Appendix A.1.

3.2 Comparison Baselines

Following the observation in AIGIBench Li et al. (2025c) that many previous detectors drop to nearly 50% accuracy under JPEG95 compression, we select recent detectors with relatively stronger robustness for evaluation, including DRCT Chen et al. (2024), Effort Yan et al. (2024), B-Free Guillaro et al. (2025), DDA Chen et al. (2025), MIRROR Liu et al. (2026), and OmniAID Guo et al. (2025). We conduct all experiments using only their official code and released pre-trained models.

3.3 Finding 1: A Pronounced Gap between Natural and Document Images

To ensure a fair comparison across detectors, we calibrate the decision threshold of each model such that its Real Acc. is fixed at 95%. Under this setting, we report the Fake Acc. and AUC of each detector on AIGDoc-Pilot in Table 1. Ideally, AI-generated image detectors should be largely independent of image semantics, as they are designed to capture generation artifacts rather than category-specific visual content. Under the same generator and generation settings, their detection performance should therefore remain comparable between AI-generated natural images and document images. However, Table 1 clearly violates this assumption. Across all evaluated detectors, the mean AUC consistently declines on document images, with an average drop of 7.15%. The degradation is also evident in F.Acc. for stronger detectors: B-Free and DDA drop by 13.81% and 13.25%, respectively. **These results demonstrate a pronounced domain gap between natural-image and document-image scenarios.** More details of AUC results are shown in Table 6 of Appendix.

This gap remains observable even for detectors whose training data already involves SDv2.1-related reconstruction or image-conditioned generation. Under the SDv2.1-VAE setting, B-Free, DDA, and OmniAID still show F.Acc. drops of 7.15%, 1.65%, and 29.36%, respectively, on document images, despite the generator being relatively close to their training distribution. In addition, when the noise strength of FLUX.1-dev reaches 0.8, the generated image is almost fully regenerated, and the original semantics are largely weakened. Nevertheless, MIRROR achieves 90.28% F.Acc. on natural images but only 55.41% on document images. This suggests that the gap cannot be simply attributed to semantic preservation in image-to-image generation; instead, document images intrinsically affect how generation artifacts are represented and perceived by detectors.

DRCT and Effort appear to be exceptions, as their document-image F.Acc. is slightly higher than that on natural images. However, their detection accuracy is extremely low in both domains, suggesting that they fail to reliably identify generated images. Moreover, their AUC still follows the same trend, decreasing by 1.42% and 15.27%, respectively.

Overall, these results demonstrate that existing detectors do not learn fully domain-independent artifact representations. Even under controlled generation settings, document images consistently introduce additional difficulty, making AI-generated document image detection a distinct and more challenging problem than conventional AI-generated natural image detection.

3.4 Finding 2: Strong Spatial Inconsistency of Artifacts.

Existing detectors commonly adopt centercrop during image preprocessing, implicitly assuming that the selected crop can provide representative artifact evidence for the entire image. However, the content of document images is often spatially sparse and unevenly distributed. Therefore, we examine whether detectable artifacts in AI-generated document images are uniformly distributed across different local regions.

Table 2: **Spatial Inconsistency of Fake-Score.** We report *Max* − *Center*, where *Max* is the largest fake score among sliding-window crops and *Center* is the standard center-crop. All scores are mapped to $[0, 1]$ using the Sigmoid function. Gray rows denote document images, while white rows denote natural images.

Method	SDv2.1-VAE	HiDream-E1-Full	FLUX.1-dev	Qwen-Image	SDv3.5-Large	Mean
B-Free	0.074	0.320	0.366	0.346	0.199	0.261
	0.187	0.402	0.436	0.367	0.394	0.357
DDA	0.066	0.263	0.215	0.163	0.120	0.165
	0.098	0.397	0.371	0.281	0.282	0.286
MIRROR	0.403	0.397	0.303	0.403	0.359	0.373
	0.392	0.463	0.525	0.470	0.478	0.465
PROBE	0.007	0.311	0.388	0.304	0.088	0.220
	0.054	0.464	0.488	0.296	0.286	0.318
OmniAID	0.177	0.164	0.164	0.210	0.209	0.185
	0.201	0.258	0.117	0.210	0.266	0.211

In particular, we investigate both AI-generated natural images and document images. For each image, we extract 25 crop windows arranged on a uniform 5×5 grid. The crop size is set to the standard input size of each detector, and the window includes the conventional center crop. Using AIGDoc-Pilot, we randomly sample 300 images for each generation setting. For each image, we compute the fake scores of all 25 crops and use the difference between the maximum score and the center-crop score to measure spatial inconsistency. As shown in Table 2, both AI-generated natural and document images exhibit uneven artifact responses, but the effect is more pronounced on document images. On average, the fake-score increases from 0.241 on natural images to 0.327 on document images, yielding an additional gap of 0.086. **This indicates that fake evidence in AI-generated document images is more localized and less spatially persistent**. In some cases, different regions of the same document image can even lead to inconsistent detector responses. For example, MIRROR shows an average increase of 0.465 on document images, which is large enough to substantially affect threshold-based decisions.

We attribute this inconsistency to the visual composition of document images. Compared with natural images with continuous textures and complex structures, document images are often dominated by homogeneous backgrounds and sparse high-frequency elements. Such sparsity makes artifact evidence highly dependent on local structures, while large background regions provide limited cues and may weaken the overall detector response. This further motivates a more detailed analysis of which local regions contribute to real-fake discrimination.

3.5 Finding 3: Text Density Enhances Authentic-Synthetic Separability.

To investigate how regional structure contributes to spatial inconsistency, we analyze the effect of text density on detector behavior using AIGDoc-Pilot. Specifically, we randomly sample 300 real

Table 3: **Document-Region Fake Acc. under Different Text Densities at 95% Real Acc.** Each generator reports F.Acc. (%). We also report the mean F.Acc. and mean AUC across generation settings. Rows under each detector correspond to non-text, low-text, medium-text, and high-text regions.

Method	Region	SDv2.1-VAE	HiDream-E1-Full	FLUX.1-dev				Qwen-Image		SDv3.5-Large		Mean	
				0.2	0.4	0.6	0.8	0.2	0.4	0.2	0.4	F.Acc.	AUC
B-Free	No	72.33	39.00	14.67	17.00	34.00	28.00	62.67	65.67	45.67	47.67	42.67	82.16
	Low	84.67	48.33	18.67	23.33	35.00	38.00	61.33	72.00	46.33	51.00	$47.87_{+5.20}$	$83.56_{+1.40}$
	Mid	86.67	48.00	21.00	27.00	47.00	53.00	63.00	74.00	53.33	58.67	$53.17_{+10.50}$	$85.02_{+2.86}$
	High	96.33	47.00	28.00	34.00	53.00	56.67	69.67	77.00	56.33	67.33	$\mathbf{58.53}_{+15.86}$	$\mathbf{86.42}_{+4.26}$
DDA	No	80.00	47.33	14.00	16.67	13.00	28.00	62.67	66.00	51.67	66.33	44.57	81.67
	Low	88.67	63.33	22.00	18.33	24.33	39.67	66.33	73.33	67.33	80.67	$54.40_{+9.83}$	$85.98_{+4.31}$
	Mid	93.00	64.33	29.33	28.00	38.33	51.67	67.67	74.33	68.00	79.67	$59.43_{+14.86}$	$87.63_{+5.96}$
	High	94.33	64.67	30.00	32.33	39.00	56.67	73.00	81.67	73.33	85.67	$\mathbf{63.07}_{+18.50}$	$\mathbf{88.46}_{+6.79}$
MIRROR	No	16.33	36.33	7.00	8.67	16.33	26.00	48.00	87.00	32.00	68.00	34.57	72.54
	Low	21.33	39.67	5.67	13.00	22.67	38.00	45.67	90.67	42.67	76.00	$39.53_{+4.96}$	$73.35_{+0.81}$
	Mid	21.67	39.00	7.67	15.00	26.67	40.33	43.33	92.67	35.67	74.00	$39.60_{+5.03}$	$73.99_{+1.45}$
	High	31.00	43.33	11.00	21.00	34.67	45.00	39.33	89.33	37.67	79.00	$\mathbf{43.13}_{+8.56}$	$\mathbf{74.06}_{+1.52}$
PROBE	No	83.00	38.33	10.33	10.00	12.00	6.00	44.33	59.00	38.00	63.00	36.40	76.09
	Low	94.33	50.00	15.67	12.00	19.00	15.00	52.33	67.00	56.33	78.00	$45.97_{+9.57}$	$79.17_{+3.08}$
	Mid	97.33	49.00	14.67	10.67	20.00	18.33	62.00	73.00	63.33	78.00	$48.63_{+12.23}$	$80.64_{+4.55}$
	High	98.67	56.67	19.33	12.67	25.67	28.33	62.67	62.67	64.00	81.67	$\mathbf{51.23}_{+14.83}$	$\mathbf{81.22}_{+5.14}$
OmniAID	No	20.33	42.00	7.00	9.33	9.00	15.00	27.33	42.33	28.67	56.00	**25.70**	**62.80**
	Low	19.00	37.33	6.00	7.00	7.00	15.00	24.00	43.33	31.00	51.67	$24.13_{-1.57}$	$61.91_{-0.89}$
	Mid	17.67	31.67	9.67	9.67	8.00	17.67	22.00	47.00	27.33	53.00	$24.37_{-1.33}$	$62.34_{-0.46}$
	High	21.33	34.67	9.67	8.33	9.00	16.67	24.67	44.33	28.00	52.67	$24.93_{-0.77}$	$61.66_{-1.14}$

document images and all their corresponding generated images. Using a lightweight proxy metric based on text-stroke density, we select four categories of cropped regions: no-, low-, medium-, and high-text-density regions. We then evaluate each detector on these region-level crops under the same 95% RAcc. protocol. More details about calculating text density are shown in Appendix A.3.

As shown in Table 3, **text density exhibits a clear positive relationship with real-synthetic separability for most detectors**. Specifically, B-Free, DDA, and MIRROR increase the F.Acc. monotonically from text-free to high-text-density regions, with gains of 15.86%, 18.50%, and 8.56%, respectively. Their AUC values also increase consistently, indicating that text-dense regions improve not only threshold-based accuracy but also the overall ranking ability between real and generated samples. OmniAID is the main exception, showing no consistent improvement with increasing text density. We conjecture that this may be related to its semantic-aware design, which makes its predictions less directly dependent on local text-density variations.

Interestingly, this observation appears to contrast with TextFake benchmark Zhang et al. (2026), which reports a 'Text Density Curse', where detection accuracy decreases as text density increases. To better understand this discrepancy, we further report the F.Acc. under the default threshold of 0.5 in Table 8 of the Appendix. Our results show a similar trend to TextFake: denser text regions can lead to lower F.Acc. for several detectors. However, this does not necessarily imply weaker real-synthetic separability. As shown in Table 3, text-dense regions achieve higher F.Acc. and AUC for most detectors. Moreover, the calibrated thresholds generally decrease as text density increases, which is shown in Table 4, indicating a clear threshold shift across text-density levels. Therefore, the apparent degradation under a fixed threshold is largely caused by threshold misalignment rather than the absence of discriminative artifacts. In other words, text density makes fixed-threshold decisions less reliable, but improves the intrinsic separability between real and generated document. More details can be found in Appendix A.3.

Table 4: **Calibrated Thresholds under Different Text-Density Regions.** For each detector, the threshold is selected to maintain 95% R.Acc. on document regions.

Method	No	Low	Mid	High
B-Free	0.779	0.513	0.342	0.214
DDA	0.483	0.384	0.344	0.309
MIRROR	0.470	0.156	0.140	0.112
PROBE	0.620	0.240	0.161	0.100
OmniAID	0.390	0.395	0.381	0.371

In conclusion, higher text density improves the separability between real and AI-generated document images. This further validates our analysis in Sec 3.4: large homogeneous background regions provide limited forensic cues and weaken detector responses, while discriminative artifacts are more likely to appear around text-related high-frequency structures.

4 AIGDoc: A Real-World Multi-Generator Benchmark for AI-Generated Document Image Detection

To enable comprehensive evaluation of AI-generated image detectors in document-centric scenarios, we introduce **AIGDoc**, a real-world, multi-generator benchmark for AI-generated document image detection. AIGDoc comprises document images collected from diverse practical settings, including camera-captured and born-digital documents, together with AI-generated counterparts produced by multiple advanced generation and editing models. It provides a systematic testbed for evaluating detection performance, cross-generator generalization, and document-specific failure modes.

4.1 Training Setting

We select 46k camera-captured document images [2] collected from real-world scenarios and reconstruct them using the SDv2.1 VAE. The image pairs are used to train our document-image detector. We adopt PE-Core-L14-336 Bolya et al. (2025) as the backbone and fine-tune it with LoRA Hu et al. (2022), setting $r = 8$ and $\alpha = 16$. During training, we use the Adam optimizer with a batch size of 64. We train three model variants as follows:

PE-LoRA. We use 336×336 center-cropped windows during both training and inference. For real images, we use MSCOCO Lin et al. (2014) and reconstruct them using the SDv2.1 VAE.

PE-LoRA†. This variant keeps the same training data and optimization settings as PE-LoRA, but replaces center cropping with the text-density-based region selection strategy described above. Specifically, we select the region with the highest text density for both training and inference.

PE-LoRA‡. This variant follows the same text-density-based training and inference strategy as PE-LoRA†, but is trained using the document training split of AIGDoc.

4.2 Evaluation Datasets

To cover diverse image generation settings, AIGDoc is organized into multiple subsets. Each subset contains an equal number of real and synthetic images. Details of the generation models and construction procedures for the evaluation subsets are provided in Appendix A.6.

Real Image Construction. We collect document images from a wide range of real-world scenarios Artaud et al. (2018); Sun et al. (2021); Cheng et al. (2023); Yu et al. (2023); Zhang et al. (2024; 2025b); Wang et al. (2026b) and manually filter them to retain high-quality camera-captured samples, which cover diverse document types such as receipts, invoices, books, newspapers, and examination papers. The collected documents also span multiple languages, including Chinese, English, and so on, providing broad coverage of practical document appearances. In addition, to enrich the data composition and enable evaluation under substantially different acquisition forms, we also include born-digital document images.

Evaluation Datasets. To better approximate real-world forgery scenarios, we consider both text-to-image generation and image editing, which correspond to different data distributions. The visualization is shown in Figure 1 (b).

i) Text-to-image Generation. To obtain fine-grained and accurate prompts for document synthesis, we first use InternVL3.5-38B Wang et al. (2025b) to generate detailed descriptions for the original document images. The model is instructed to produce structured outputs containing three aspects: *Overall Description*, *Layout Structure*, and *Image Text*. We further refine these prompts with InternVL3.5-38B to enrich prompt diversity while removing potentially sensitive information. Based on these, we generate document images using a set of recent text-to-image models, including HiDream-O1-Image Cai et al. (2026), Z-Image-Turbo Cai et al. (2025a), GPT-Image-2 OpenAI (2026), Nano-Banana-2 Google (2026), and Seedream-4.5 ByteDance Seed (2025). Each subset contains 1,000 real images and 1,000 AI-generated images. Representative prompt examples and their corresponding generated images are shown in Fig. 6 in the Appendix.

ii) Image Editing. To simulate realistic document forgery scenarios, we further construct AI-generated document images through image editing. Unlike text-to-image generation, image editing

[2] All training data are kept strictly disjoint from the real images used in the test dataset.

Table 5: **Detection Accuracy across Different AI-Generated Document Image Settings.** Each detector reports Acc. (%) for each generation setting.

Generator	UnivFD	DRCT	Effort	B-Free	DDA	MIRROR	OmniAID	PROBE	SICA	PE-LoRA	PE-LoRA†	PE-LoRA‡
Text2Image												
Z-Image-Turbo	49.4	61.8	85.2	52.3	61.4	49.9	49.8	55.1	44.9	68.0	**92.8**	90.5
HiDream-O1-Image	56.5	62.1	72.7	97.0	95.2	66.8	63.2	**97.9**	75.8	95.2	**97.9**	93.5
Seedream 4.5	50.1	51.7	58.2	95.2	91.8	52.8	53.3	**99.5**	61.8	98.0	97.4	95.7
Nano-Banana-2	50.0	46.2	60.2	66.0	58.9	50.4	51.0	63.9	55.4	59.8	**67.5**	60.5
GPT-Image-2	50.0	49.5	74.2	51.3	47.0	49.8	50.3	51.2	47.4	49.4	59.7	**83.9**
Image-Edit												
FLUX.1-Kontext	49.8	51.2	54.5	49.6	53.7	50.0	51.2	50.7	51.9	51.5	58.9	**93.5**
FLUX.2-Klein	49.5	43.7	48.3	50.7	52.1	50.0	50.5	51.1	52.0	50.2	53.3	**83.9**
Hunyuan-3-Instruct	49.8	57.0	61.1	72.4	95.8	54.2	66.0	96.6	64.5	66.4	97.0	**99.5**
LongCat-Image-Edit	49.8	47.9	53.3	71.9	68.0	50.4	53.1	69.1	56.8	75.7	84.2	**99.2**
OmniGen2	56.1	57.2	58.1	81.3	65.7	72.4	66.9	73.8	68.5	72.2	83.4	**94.1**
Qwen-Image-Edit	51.0	49.8	57.2	82.2	78.9	53.7	56.4	81.8	60.4	73.3	89.8	**97.3**
Step1X-Edit	49.6	55.5	59.0	74.3	76.2	50.6	53.7	78.5	55.5	75.1	82.3	**84.5**
Seedream 4.5	51.4	62.1	54.8	97.2	94.7	50.2	53.0	99.2	53.5	97.1	99.2	**99.5**
Seedream 5.0-Lite	52.8	62.1	48.3	95.6	85.8	50.3	52.7	**98.9**	54.4	93.6	98.0	98.1
Nano-Banana-2	49.3	45.0	54.9	49.0	50.8	50.0	54.4	50.4	51.3	50.2	57.4	**60.6**
Digital-Image-Edit												
Qwen-Image-Edit	47.8	34.9	50.1	30.6	23.9	50.0	50.0	39.1	8.3	55.9	**75.4**	65.6
Nano-Banana-2	48.1	36.8	24.5	17.5	24.6	**51.1**	50.0	36.1	50.9	49.5	50.1	50.0
Average	50.6	51.4	57.3	66.7	66.1	53.1	54.4	70.2	53.7	69.5	79.1	**85.3**

preserves the original document layout, paper texture, camera-captured degradation, and background appearance, while modifying only localized sensitive content. Specifically, we use InternVL3.5-38B Wang et al. (2025b) to identify candidate text regions and generate editing instructions for critical fields, such as amounts, years, dates, and key terms. Based on these localized prompts, we apply recent image editing models, including FLUX.1-Kontext Black Forest Labs et al. (2025), FLUX.2-Klein, HunyuanImage-3-Instruct Cao et al. (2025), Longcat-Image-Edit Team et al. (2025), OmniGen2 Wu et al. (2025b), Qwen-Image-Edit Wu et al. (2025a), Step1X-Edit Liu et al. (2025b), Seedream 4.5, Seedream 5.0-Lite ByteDance Seed (2026) and Nano-Banan-2. This setting reflects a practical manipulation scenario where forged document images remain visually close to the original real images but contain altered semantic information. Each subset contains 2000 real images and their paired AI-generated counterparts.

iii) Born-digital image editing. Born-digital document images also commonly appear in real-world document workflows. To better cover this practical scenario, we additionally construct edited samples from born-digital documents. Specifically, we apply Qwen-Image-Edit and Nano Banana 2 to modify sensitive textual fields in electronic document images, providing a complementary setting to camera-captured document editing. Each subset contains 1000 real images and their paired AI-generated counterparts.

4.3 Evaluation Baselines

To comprehensively evaluate the detection performance of existing detectors on document images, we selected 9 state-of-the-art detectors, including UnivFD Ojha et al. (2023), DRCT Chen et al. (2024), Effort Yan et al. (2024), B-Free Guillaro et al. (2025), DDA Chen et al. (2025), MIRROR Liu et al. (2026), OmniAID Guo et al. (2025), PROBE Cao et al. (2026), and SICA Du et al. (2026).

4.4 Detection Performance

We report the accuracy (Acc.) results on the AIGDoc dataset in Table 5. Although existing detectors demonstrate strong generalization on natural-image benchmarks, their performance degrades substantially on AI-generated document images. Most methods achieve an mAcc. below 70%, indicating that document-centric scenarios remain challenging for general-purpose detectors. Among the evaluated methods, PROBE achieves the highest mAcc. of 70.2%. Notably, compared with PE-LoRA, incorporating text-density-guided region selection improves mAcc. by 9.6%, which is consistent with our diagnostic findings: discriminative evidence in AI-generated document images tends to concentrate in text-dense, high-frequency regions, whereas homogeneous background regions provide limited forensic cues. In addition, incorporating document images during training partially narrows the natural-document gap, improving detection accuracy by 6.2% over PE-LoRA†.

Furthermore, SICA, which unifies Deepfake, AI-generated image, image-manipulation, and document-forgery detection within a single model, achieves an mAcc. of only 53.7% on AIGDoc. Despite being designed to cover document forgery and other forensic domains, it still struggles to detect AI-generated document images. This further confirms a significant gap between AI-generated document detection and the fields of document tampering detection and existing AI-generation detection methods, which also underscores that AI-generated document detection is not simply addressed by combining conventional document-forgery detection with general AI-generated image detection, highlighting its distinct research challenges.

Moreover, born-digital image editing is the most challenging setting, where many detectors exhibit near-chance or severely degraded performance. Specifically, under Qwen-Image-Edit, the performance of DDA and PROBE decreases by 55.0% and 42.7%, respectively, falling far below the chance level. These results suggest that detectors trained primarily on camera-captured images may fail to generalize to digital document images, highlighting the need for broader document-centric training and evaluation protocols. The corresponding AUC results are shown in Appendix A.5.

5 Conclusions, Limitations, and Broader Impacts

Conclusions. Our study shows that AI-generated document image detection remains substantially more challenging than conventional natural-image detection. Under controlled generation settings, AIGDoc-Pilot reveals a pronounced performance gap between natural and document images. Further analysis indicates that generation artifacts in documents are spatially inconsistent and that text-dense regions provide stronger evidence for real-synthetic discrimination. The broader evaluation on AIGDoc confirms that existing detectors remain unreliable across diverse document generation and editing settings. Although text-density-guided region selection and document-based training improve performance, they do not fully resolve this challenge.

Limitations. Despite the improvements achieved in this work, reliable detection of AI-generated document images remains unresolved. Our controlled analysis primarily focuses on camera-captured documents, whereas born-digital document editing has proven substantially more challenging for existing detectors. Future work will expand the benchmark to cover a broader range of difficult scenarios, including user interfaces, screenshots, reproduced document images, and other challenging acquisition settings.

Broader Impacts. This work reveals the practical limitations of current AI-generated image detectors in document-centric scenarios and identifies the underlying causes of their performance gap relative to natural images. We hope that AIGDoc will foster document-aware evaluation protocols and advance AI-generated image detection beyond conventional natural-image benchmarks toward more reliable real-world use. Nonetheless, a potential ethical concern is that malicious actors may exploit AIGDoc to develop more evasive AI-generated document methods.

Ethics Statement

AIGDoc is constructed for research on AI-generated document image detection. We review the licenses of all source datasets and the terms of generation APIs, and release only samples permitted

for redistribution and derivative use. We also remove sensitive personal information and prohibit the use of AIGDoc for deceptive document generation.

AI USE STATEMENT

In this work, we used generative AI tools for generating text prompts of image generation and generating AI-generated document images. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Chloé Artaud, Nicolas Sidere, Antoine Doucet, Jean-Marc Ogier, and Vincent Poulain D'Andecy Yooz. Find it! fraud detection contest report. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 13–18. IEEE, 2018.

Manoela W Auad, Sarah C Alves, Gabriel S Kakizaki, Julio C. S. Reis, and Michel M Silva. A filtering and image preparation approach to enhance ocr for fiscal receipts. In *Proceedings of the 37th Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2024. URL http://urlib.net/ibi/8JMKD3MGPEW34M/4C22QAL.

Black Forest Labs. Flux.1-dev. https://huggingface.co/black-forest-labs/FLUX.1-dev. 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Daniel Li, Piotr Dollár, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv:2504.13181*, 2025.

ByteDance Seed. Seedream 4.5. https://seed.bytedance.com/en/seedream4_5, 2025. Official model page. Accessed: 2026-07-12.

ByteDance Seed. Seedream 5.0 lite. https://seed.bytedance.com/en/seedream5_0_lite, 2026. Official model page. Accessed: 2026-08-27.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Xin Jin, Liangchen Li, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025a.

Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, et al. Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. *arXiv preprint arXiv:2505.22705*, 2025b.

Qi Cai, Jingwen Chen, Chengmin Gao, Zijian Gong, Yehao Li, Yingwei Pan, Yi Peng, Zhaofan Qiu, Kai Yu, Yiheng Zhang, et al. Hidream-o1-image: A natively unified image generative foundation model with pixel-level unified transformer. *arXiv preprint arXiv:2605.11061*, 2026.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025.

Zijie Cao, Weijie Tu, Yao Xiao, Weijian Deng, Liang Lin, and Pengxu Wei. Where detectors fail: Probing generative space for generalizable ai-generated image detection. *arXiv preprint arXiv:2605.24906*, 2026.

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

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.

Zehao Chen and Hua Yang. Attentive semantic exploring for manipulated face detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1985–1989. IEEE, 2021.

Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, and Lianwen Jin. M6doc: A large-scale multi-format, multi-type, multi-layout, multi-language, multi-annotation category dataset for modern document layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15138–15147, June 2023.

Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12830–12839, 2025.

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

Bo Du, Xiaochen Ma, Xuekang Zhu, Zhe Yang, Chaogun Niu, Chenfan Qu, Mingqi Fang, Zhenming Wang, Jingjing Liu, Jian Liu, et al. Can we build a monolithic model for fake image detection? sica: Semantic-induced constrained adaptation for unified-yet-discriminative artifact feature space reconstruction. *arXiv preprint arXiv:2602.06676*, 2026.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Google. Nano Banana 2 (Gemini 3.1 Flash Image). `https://ai.google.dev/gemini-api/docs/image-generation`, 2026. Gemini API image generation documentation. Model ID: gemini-3.1-flash-image. Accessed: 2026-07-12.

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.

Yuncheng Guo, Junyan Ye, Chenjue Zhang, Hengrui Kang, Haohuan Fu, Conghui He, and Weijia Li. Omniaid: Decoupling semantic and artifacts for universal ai-generated image detection in the wild. *arXiv preprint arXiv:2511.08423*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Xinjue Hu, Zhangjie Fu, Xiang Zhang, and Yanyu Chen. Invisible and steganalysis-resistant deep image hiding based on one-way adversarial invertible networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7):6128–6143, 2023.

Xinjue Hu, Zhenshan Tan, Chi Wang, Xiang Zhang, and Zhangjie Fu. High-resolution image steganalysis. *IEEE Transactions on Dependable and Secure Computing*, 2026.

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

Manyi Li, Renshuai Tao, Yufan Liu, Chuangchuang Tan, Haotong Qin, Bing Li, Yunchao Wei, and Yao Zhao. Pay less attention to deceptive artifacts: Robust detection of compressed deepfakes on online social networks. *arXiv preprint arXiv:2506.20548*, 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.

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? In *Advances in Neural Information Processing Systems*, 2025c.

Kaiqing Lin, Yuzhen Lin, Weixiang Li, Taiping Yao, and Bin Li. Standing on the shoulders of giants: Reprogramming visual-language model for general deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5262–5270, 2025.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

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*, 2025a.

Ruiqi Liu, Manni Cui, Ziheng Qin, Zhiyuan Yan, Ruoxin Chen, Yi Han, Zhiheng Li, Junkai Chen, ZhiJin Chen, Kaiqing Lin, et al. Mirror: Manifold ideal reference reconstructor for generalizable ai-generated image detection. *arXiv preprint arXiv:2602.02222*, 2026.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025b.

Zihan Liu, Hanyi Wang, Yaoyu Kang, and Shilin Wang. Mixture of low-rank experts for transferable ai-generated image detection. *arXiv preprint arXiv:2404.04883*, 2024.

Dongliang Luo, Yuliang Liu, Rui Yang, Xianjin Liu, Jishen Zeng, Yu Zhou, and Xiang Bai. Toward real text manipulation detection: New dataset and new solution. *Pattern Recognition*, 157:110828, 2025.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16317–16326, 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.

OpenAI. GPT Image 2. `https://developers.openai.com/api/docs/models/gpt-image-2`, 2026. OpenAI API model documentation. Accessed: 2026-07-12.

Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Omni-iml: Towards unified interpretable image manipulation localization. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Chenfan Qu, Yiwu Zhong, Jian Liu, Xuekang Zhu, Bohan Yu, and Lianwen Jin. Textshield-r1: Reinforced reasoning for tampered text detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8621–8629, 2026b.

Chenfan Qu, Yiwu Zhong, Xuekang Zhu, Junchi Li, Changjiang Jiang, Lianwen Jin, et al. Detect any ai-counterfeited text image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35437–35450, 2026c.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Yalin Song, Wenbin Jiang, Xiuli Chai, Zhihua Gan, Mengyuan Zhou, and Lei Chen. Cross-attention based two-branch networks for document image forgery localization in the metaverse. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(2):1–24, 2025.

Hongbin Sun, Zhanghui Kuang, Xiaoyu Yue, Chenhao Lin, and Wayne Zhang. Spatial dual-modality graph reasoning for key information extraction. *arXiv preprint arXiv:2103.14470*, 2021.

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.

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.

Meituan LongCat Team, Hanghang Ma, Haoxian Tan, Jiale Huang, Junqiang Wu, Jun-Yan He, Lishuai Gao, Songlin Xiao, Xiaoming Wei, Xiaoqi Ma, et al. Longcat-image technical report. *arXiv preprint arXiv:2512.07584*, 2025.

Chi Wang, Ziwen He, Xinjue Hu, Weinan Guan, Wei Wang, and Zhangjie Fu. Map-mamba: Multi-artifacts perception mamba for generalizable face forgery detection. *IEEE Transactions on Information Forensics and Security*, 2026a.

Fan Wang, Xiang Zhang, and Zhangjie Fu. An iterative two-stage probability adjustment strategy with progressive incremental searching for image steganography. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9428–9444, 2024.

Fan Wang, Zhangjie Fu, Xiang Zhang, Ziqiang Li, Ziwen He, and Manyu Wang. Pair-wise confidence difference-based pseudo-label selection for universal mismatched steganalysis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 8673–8681, 2025a.

Hang Wang, Jin Zhang, Guoliang Xu, Pengyue Lu, Yao Li, Zijiao Zhang, Tianyu Huang, Weiqi Xiong, Yulong Wang, Chuqiao Lu, et al. Finixdoc: Rethinking financial document parsing beyond saturated benchmarks. 2026b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

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.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, Ze Liu, Ziyi Xia, Chaofan Li, Haoge Deng, Jiahao Wang, Kun Luo, Bo Zhang, Defu Lian, Xinlong Wang, Zhongyuan Wang, Tiejun Huang, and Zheng Liu. Omnigen2: Towards instruction-aligned multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Yi Wu, Ziqiang Li, Heliang Zheng, Chaoyue Wang, and Bin Li. Infinite-id: Identity-preserved personalization via id-semantics decoupling paradigm. In *European Conference on Computer Vision*, pp. 279–296. Springer, 2024.

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.

Jiazhen Yan, Fan Wang, Weiwei Jiang, Ziqiang Li, and Zhangjie Fu. Ns-net: Decoupling clip semantic information through null-space for generalizable ai-generated image detection. *arXiv preprint arXiv:2508.01248*, 2025b.

Jiazhen Yan, Ziqiang Li, Fan Wang, Ziwen He, and Zhangjie Fu. Dual frequency branch framework with reconstructed sliding windows attention for ai-generated image detection. *IEEE Transactions on Information Forensics and Security*, 2026.

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.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Wenwen Yu, Chengquan Zhang, Haoyu Cao, Wei Hua, Bohan Li, Huang Chen, Mingyu Liu, Mingrui Chen, Jianfeng Kuang, Mengjun Cheng, et al. Icdar 2023 competition on structured text extraction from visually-rich document images. In *International Conference on Document Analysis and Recognition*, pp. 536–552. Springer, 2023.

Haifeng Zhang, Qinghui He, Xiuli Bi, Weisheng Li, Bo Liu, and Bin Xiao. Towards universal ai-generated image detection by variational information bottleneck network. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23828–23837, 2025a.

Shuo Zhang, Biao Yang, Zhang Li, Zhiyin Ma, Yuliang Liu, and Xiang Bai. Exploring the capabilities of large multimodal models on dense text. In *International Conference on Document Analysis and Recognition*, pp. 281–298. Springer, 2024.

Weiguang Zhang, Huangcheng Lu, Maizhen Ning, Xiaowei Huang, Wei Wang, Kaizhu Huang, and Qiufeng Wang. Dvd: Unleashing a generative paradigm for document dewarping via coordinates-based diffusion model. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025b.

Yuning Zhang, Changtao Miao, Mingyu Liao, Tingyu Liu, Xinghao Wang, Tao Gong, Qi Chu, and Nenghai Yu. Textfake: Benchmarking ai-generated image detection on text-rich images. *arXiv preprint arXiv:2606.01050*, 2026.

Yue Zhou, Xinan He, KaiQing Lin, Bin Fan, Feng Ding, and Bin Li. Breaking latent prior bias in detectors for generalizable aigc image detection. *arXiv preprint arXiv:2506.00874*, 2025a.

Yue Zhou, Xinan He, Kaiqing Lin, Bing Fan, Feng Ding, Jinhua Zeng, and Bin Li. Brought a gun to a knife fight: Modern vfm baselines outgun specialized detectors on in-the-wild ai image detection. *arXiv preprint arXiv:2509.12995*, 2025b.

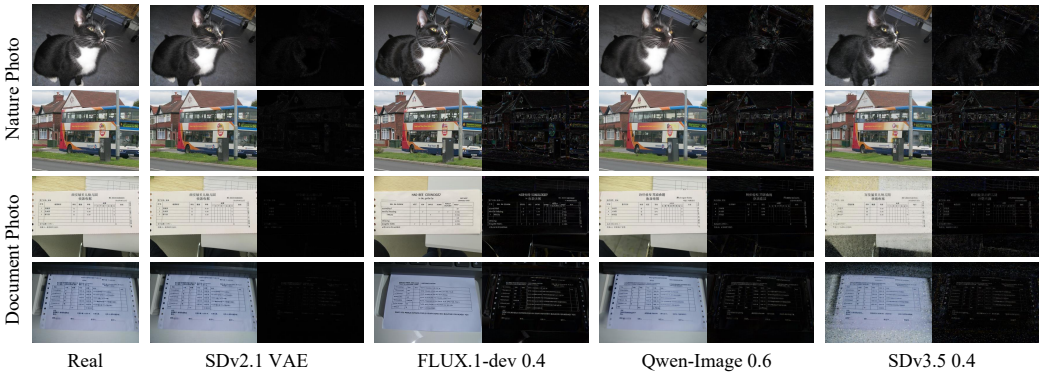


Figure 2: **Qualitative Visualization of AIGDoc-Pilot Generation Results.** For each real image, we show the generated counterpart and its difference map. Compared with natural images, document images exhibit much sparser generation-induced differences, which are mainly concentrated around text, table lines, and other high-frequency structures.

Table 6: **Detection Performance Gap between AI-Generated Natural and Document Images on AIGDoc-Pilot.** Each generator setting reports AUC (%). Gray rows denote results on document images, while white rows denote results on natural images.

Method	SDv2.1-VAE	HiDream-E1-Full	FLUX.1-dev				Qwen-Image		SDv3.5-Large		Mean AUC
			0.2	0.4	0.6	0.8	0.2	0.4	0.2	0.4	
DRCT	69.25	69.79	56.93	65.91	73.19	81.69	70.52	74.90	65.13	70.52	**69.78**
	67.88	57.76	61.23	58.91	66.63	79.09	75.30	80.94	67.28	68.54	68.36$_{-1.42}$
Effort	48.80	49.02	41.53	55.24	64.54	72.91	61.74	64.30	47.95	51.31	**55.73**
	30.70	27.26	29.86	36.48	48.73	64.40	57.17	59.23	24.46	26.32	40.46$_{-15.27}$
B-Free	99.63	90.46	74.47	70.94	74.83	83.55	94.24	95.68	98.14	98.80	**88.08**
	98.01	85.08	64.55	66.13	79.96	82.81	90.83	93.66	89.08	92.17	84.23$_{-3.85}$
DDA	99.31	91.12	82.24	85.71	89.70	89.71	96.22	96.31	97.49	97.62	**92.54**
	98.67	83.59	70.41	71.40	77.12	87.32	93.18	96.61	93.11	96.06	86.75$_{-5.80}$
MIRROR	77.82	74.13	58.02	80.82	90.37	96.63	83.66	87.70	77.83	90.81	**81.78**
	74.48	70.90	56.42	67.80	81.59	86.44	81.44	98.34	70.14	92.98	78.05$_{-3.73}$
OmniAID	84.26	82.57	54.30	62.49	68.77	85.57	72.81	77.52	73.39	86.51	**74.82**
	61.73	76.25	52.53	37.38	35.90	51.57	67.98	80.09	72.66	84.19	62.03$_{-12.79}$

A Appendix

A.1 More details of AIGDoc-Pilot

The AIGDoc-Pilot dataset contains two subsets: a natural-image subset and a document-image subset. Both subsets consist of camera-captured real images, with 545 images in each category. For each real image, we construct the corresponding synthetic samples using image-to-image generation or VAE reconstruction, while keeping generation settings identical for natural and document images.

Figure 2 provides qualitative examples of the generated images and their pixel-level differences from the corresponding real images. For natural images, the difference maps usually exhibit spatially continuous responses over objects, textures, and backgrounds, indicating that generation-induced changes are distributed across broad visual regions. In contrast, the difference maps of document images are much sparser and mainly concentrate around text characters, table lines, stamps, and other high-frequency structures, while large homogeneous background areas contain only weak changes, highlighting the difficulty of detecting AI-generated document images.

A.2 More Results and Analysis of Finding 1

We additionally report the AUC of each detector across different generator settings in Table 6. All detectors exhibit lower AUC on document images than on natural images, with an average decrease

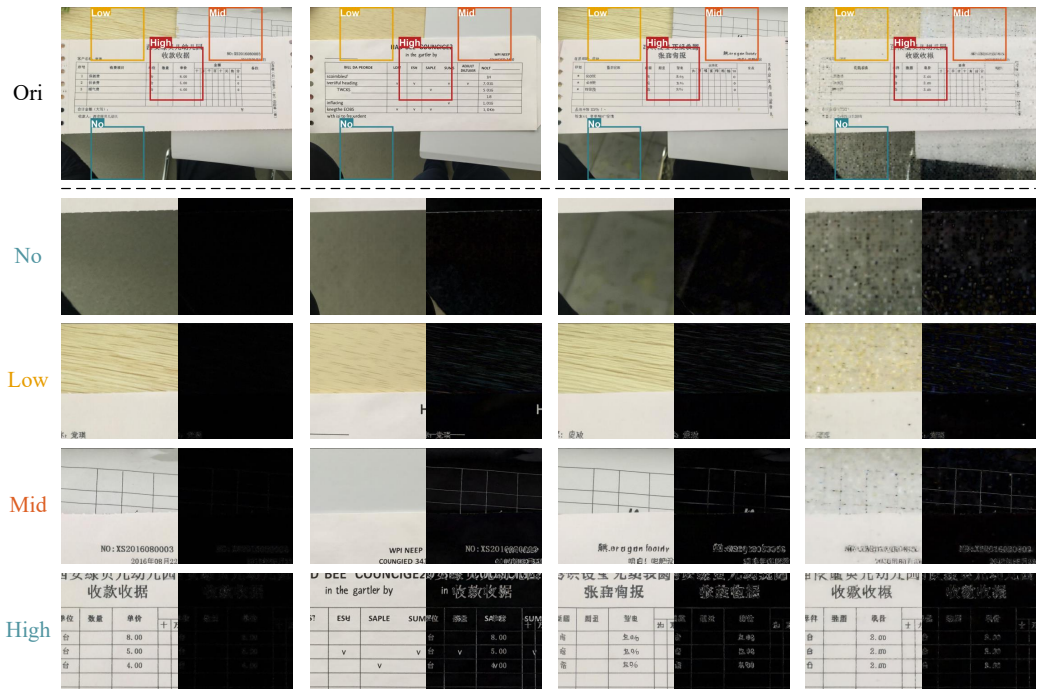


Figure 3: **Visualization of text-density-based crop regions and their corresponding residual maps.** Low-text-density regions exhibit weaker and less discernible differences between generated and real images, whereas text-dense regions reveal more pronounced generation traces.

of 7.15%, ranging from 1.42% for DRCT to 15.27% for Effort. This consistent degradation further reinforces Finding 1: A pronounced performance gap exists between natural and document images.

A.3 More Results and Analysis of Finding 3

To examine the effect of text density on detection performance, we develop a region-selection algorithm that identifies four regions with distinct text densities in each document image. Specifically, we sample candidate 448×448 patches from each real image and compute the proportion of dark, high-gradient pixels within each patch, which serves as a lightweight proxy for dense text strokes and boundaries. Based on this score, we select spatially distinct no-, low-, medium-, and high-text-density regions while limiting their mutual overlap. The selected coordinates are determined solely from the real image and identically applied to its paired generated image, ensuring spatially aligned regional comparisons. Each selected patch is subsequently center-cropped to meet the input requirements of the corresponding detector. As shown in Fig. 3, the selected regions exhibit clearly distinct text densities. Their residual maps further show that text-dense regions tend to exhibit more pronounced localized differences between real and generated images, suggesting that they may provide more discriminative artifact evidence.

In addition, we report AUC to assess detector performance independently of a specific decision threshold. As shown in Table 7, the AUC of most detectors consistently increases with text density, indicating that text-dense regions provide stronger discriminative cues for separating real and AI-generated document images.

However, when using the fixed threshold of 0.5 commonly adopted in existing evaluation protocols, we observe an opposite trend in Table 8: higher text density leads to lower F.Acc., which appears to suggest increased detection difficulty. This discrepancy arises from a score distribution shift across text-density levels. As text density increases, the overall detector scores for both real and generated regions tend to decrease, causing the calibrated threshold for maintaining 95% Real Acc. to move downward. Therefore, a fixed threshold of 0.5 becomes increasingly misaligned with the actual

Table 7: **Document-Region AUC under Different Text Densities.** Each generator reports AUC (%). Rows under each detector correspond to non-text, low-text, medium-text, and high-text regions.

Method	Region	SDv2.1-VAE	HiDream-E1-Full	FLUX.1-dev				Qwen-Image		SDv3.5-Large		Mean
				0.2	0.4	0.6	0.8	0.2	0.4	0.2	0.4	
B-Free	No	94.03	82.70	66.38	67.15	78.94	80.44	90.23	91.21	83.84	86.64	82.16
	Low	96.63	82.89	67.51	68.90	79.69	84.52	90.09	93.23	83.92	88.26	$83.56_{+1.40}$
	Mid	96.76	83.63	68.63	71.98	83.36	85.43	91.09	93.32	86.24	89.73	$85.02_{+2.86}$
	High	98.40	81.83	70.48	74.92	84.84	88.03	93.00	94.80	87.00	90.84	$\mathbf{86.42}_{+4.26}$
DDA	No	95.78	85.87	67.37	64.15	68.58	75.90	89.94	91.51	86.22	91.34	81.67
	Low	97.88	88.59	72.88	71.67	73.89	81.94	91.90	93.97	91.88	95.20	$85.98_{+4.31}$
	Mid	98.37	89.45	76.08	74.65	76.44	86.66	92.39	94.62	92.38	95.27	$87.63_{+5.96}$
	High	98.08	88.70	74.27	76.41	80.11	87.13	93.81	95.54	94.09	96.50	$\mathbf{88.46}_{+6.79}$
MIRROR	No	59.09	70.92	54.22	59.42	73.81	80.43	75.76	95.16	67.69	88.89	72.54
	Low	62.00	71.51	53.09	60.87	74.45	80.72	72.64	96.04	71.18	90.97	$73.35_{+0.81}$
	Mid	62.37	71.89	55.51	61.03	75.66	83.26	73.56	97.59	68.72	90.29	$73.99_{+1.45}$
	High	64.42	72.88	55.84	61.81	75.48	82.47	72.32	95.36	67.99	92.05	$\mathbf{74.06}_{+1.52}$
PROBE	No	97.17	84.74	61.92	53.81	56.19	56.80	86.82	90.93	82.17	90.33	76.09
	Low	98.60	85.85	66.72	56.29	58.76	63.66	88.41	92.39	86.59	94.39	$79.17_{+3.08}$
	Mid	99.22	87.57	67.89	59.74	61.81	66.88	88.43	92.42	87.98	94.43	$80.64_{+4.55}$
	High	99.14	85.12	68.09	59.06	63.55	72.10	88.60	92.82	88.25	95.51	$\mathbf{81.22}_{+5.14}$
OmniAID	No	68.31	73.02	53.70	45.51	43.18	48.44	66.62	70.71	75.10	83.44	**62.80**
	Low	69.77	69.55	52.52	44.10	39.78	47.98	64.62	72.11	74.54	84.12	$61.91_{-0.89}$
	Mid	70.47	69.74	54.08	45.19	41.69	47.84	65.13	73.59	72.72	82.93	$62.34_{-0.46}$
	High	69.19	68.64	54.11	45.00	41.66	46.93	64.15	72.80	71.64	82.43	$61.66_{-1.14}$

Table 8: **Document-Region Fake Acc. under Different Text Densities with the 0.5 Threshold.** Each generator reports F.Acc. (%). We also report the mean F.Acc. and mean AUC across generation settings. Rows under each detector correspond to non-text, low-text, medium-text, and high-text regions.

Method	Region	SDv2.1-VAE	HiDream-E1-Full	FLUX.1-dev				Qwen-Image		SDv3.5-Large		Mean	
				0.2	0.4	0.6	0.8	0.2	0.4	0.2	0.4	F.Acc.	AUC
B-Free	No	86.33	54.67	29.00	29.67	44.33	49.33	66.67	70.00	60.00	60.33	**55.03**	82.16
	Low	85.00	45.67	20.00	21.33	37.00	43.33	56.33	69.33	46.67	50.33	$47.50_{-7.53}$	$83.56_{+1.40}$
	Mid	80.00	39.67	17.33	22.67	36.00	46.67	55.33	65.00	48.67	49.67	$46.10_{-8.93}$	$85.02_{+2.86}$
	High	82.00	38.00	18.00	22.33	33.00	42.00	56.00	65.33	42.33	49.33	$44.83_{-10.20}$	$\mathbf{86.42}_{+4.26}$
DDA	No	83.00	52.67	20.00	17.33	17.33	26.00	49.33	52.00	59.00	66.00	44.27	81.67
	Low	84.67	49.33	15.00	13.67	19.67	32.00	47.67	63.33	62.33	70.00	$45.77_{+1.50}$	$85.98_{+4.31}$
	Mid	85.33	49.67	15.33	18.00	25.00	37.67	51.33	62.67	57.67	72.00	$47.47_{+3.20}$	$87.63_{+5.96}$
	High	88.67	48.67	18.00	16.00	23.67	38.33	57.00	67.33	62.67	74.33	$\mathbf{49.47}_{+5.20}$	$\mathbf{88.46}_{+6.79}$
MIRROR	No	17.33	38.33	8.00	10.00	17.67	29.67	27.33	67.00	34.00	66.00	**31.53**	72.54
	Low	13.00	30.67	2.67	6.33	14.33	24.67	22.33	74.33	30.33	66.67	$28.53_{-3.00}$	$73.35_{+0.81}$
	Mid	12.33	29.67	2.00	5.67	14.67	22.67	22.67	76.33	29.00	63.33	$27.83_{-3.70}$	$73.99_{+1.45}$
	High	12.33	27.67	2.00	4.00	10.67	18.67	22.33	72.33	24.33	64.67	$25.90_{-5.63}$	$\mathbf{74.06}_{+1.52}$
PROBE	No	89.33	45.33	13.33	11.33	12.33	10.67	45.33	59.67	46.67	60.67	**39.47**	76.09
	Low	92.67	39.33	9.33	6.33	9.00	8.00	46.00	57.33	44.00	57.67	$36.97_{-2.50}$	$79.17_{+3.08}$
	Mid	93.00	36.33	9.00	8.33	10.00	9.67	43.33	59.00	41.33	56.67	$36.67_{-2.80}$	$80.64_{+4.55}$
	High	93.00	34.33	9.33	6.00	11.67	11.33	43.33	48.67	39.33	54.67	$35.17_{-4.30}$	$\mathbf{81.22}_{+5.14}$
OmniAID	No	12.67	31.33	2.67	5.00	5.00	9.33	19.33	30.33	20.00	42.00	**17.77**	**62.80**
	Low	11.67	29.00	2.33	3.00	5.00	11.67	17.00	34.33	19.67	41.67	$17.53_{-0.24}$	$61.91_{-0.89}$
	Mid	10.33	29.00	3.33	4.00	4.33	11.33	16.67	33.33	17.00	41.33	$17.07_{-0.70}$	$62.34_{-0.46}$
	High	12.33	29.67	3.33	4.00	5.33	9.00	18.00	33.67	16.33	42.67	$17.43_{-0.34}$	$61.66_{-1.14}$

decision boundary, producing an apparent performance degradation despite improved real-synthetic separability. This is further verified by Table 4, where most detectors require progressively lower thresholds as text density increases.

A.4 Details of Compared Detectors

To provide a comprehensive benchmark, we evaluate 9 existing state-of-the-art detection methods, including UnivFD Ojha et al. (2023), DRCT Chen et al. (2024), Effort Yan et al. (2024), B-Free Guillaro et al. (2025), DDA Chen et al. (2025), MIRROR Liu et al. (2026), OmniAID Guo et al. (2025),

Table 9: **Detection AUC across Different AI-Generated Document Image Settings.** Each detector reports AUC (%) for each generation setting.

Generator	UnivFD	DRCT	Effort	B-Free	DDA	MIRROR	OmniAID	PROBE	SICA	PE-LoRA	PE-LoRA†	PE-LoRA‡
Text2Image												
Z-Image-Turbo	48.0	69.5	93.2	60.9	78.4	56.4	31.8	86.1	42.9	84.2	**97.7**	97.2
HiDream-O1-Image	82.6	66.7	83.5	99.6	99.1	93.5	87.1	**99.9**	78.9	99.3	**99.9**	98.7
Seedream 4.5	67.8	51.4	64.0	99.0	97.8	67.2	41.0	**100.0**	64.8	99.6	99.9	99.8
Nano-Banana-2	55.9	45.2	63.6	70.6	63.6	67.8	47.0	78.5	57.5	58.2	**79.0**	64.5
GPT-Image-2	67.7	51.0	83.3	47.3	46.9	64.4	38.2	75.8	48.3	35.8	72.7	**94.3**
Image-Edit												
FLUX.1-Kontext	46.8	53.2	58.5	47.6	72.4	54.3	54.5	71.4	58.9	61.3	68.6	**99.4**
FLUX.2-Klein	43.5	42.4	36.9	44.9	54.9	50.3	54.0	51.3	58.4	41.7	41.2	**91.8**
Hunyuan-3-Instruct	53.3	62.1	72.6	87.3	99.4	74.8	66.8	99.7	78.7	86.6	99.7	**100.0**
LongCat-Image-Edit	44.1	49.0	55.9	87.4	87.1	61.2	58.8	89.5	70.5	94.6	96.6	**100.0**
OmniGen2	68.8	62.9	67.2	85.4	75.2	82.7	70.6	90.3	76.9	92.1	90.3	**99.3**
Qwen-Image-Edit	55.9	51.2	64.8	92.1	94.0	68.8	72.0	96.4	70.6	89.7	97.0	**99.8**
Step1X-Edit	42.1	58.0	66.7	80.0	84.9	58.4	61.3	88.0	67.7	85.9	83.0	**97.0**
Seedream 4.5	56.3	69.1	68.5	99.4	99.4	59.5	52.4	**100.0**	64.0	99.8	**100.0**	**100.0**
Seedream 5.0-Lite	62.1	68.1	58.1	98.7	97.4	57.7	52.2	**99.9**	63.7	99.2	99.8	99.8
Nano-Banana-2	32.9	44.2	67.0	48.1	63.4	60.7	57.7	74.8	56.4	56.2	**86.6**	85.5
Digital-Image-Edit												
Qwen-Image-Edit	36.9	29.5	48.1	19.4	14.2	6.3	10.2	2.0	2.5	77.0	91.2	**99.2**
Nano-Banana-2	3.8	30.0	14.9	15.6	13.3	**61.3**	54.0	0.0	57.0	11.2	28.9	13.3
Average	51.1	53.1	62.8	69.6	73.0	61.5	53.5	76.7	59.9	74.8	84.2	**90.6**

PROBE Cao et al. (2026), and SICA Du et al. (2026). The compared detectors are summarized as follows.

1) UnivFD (CVPR 2023) Ojha et al. (2023). UnivFD investigates the generalization ability of large-scale vision-language representations for fake image detection. It uses CLIP features with simple nearest-neighbor or linear-probing classifiers, showing that pre-trained feature spaces can provide strong cross-generator detection capability without training a detector from scratch.

2) DRCT (ICML 2024) Chen et al. (2024). DRCT introduces diffusion reconstruction contrastive training for universal detection of diffusion-generated images. It reconstructs images using diffusion models to generate hard training samples and leverages contrastive learning to encourage the detector to capture diffusion-related artifacts with improved generalization.

3) Effort (ICML 2025) Yan et al. (2024). Effort observes that naively fine-tuned detectors can overfit to limited fake patterns, leading to a low-rank and poorly generalizable feature space. It decomposes the feature space into orthogonal subspaces via SVD, preserving the principal pre-trained knowledge while adapting residual components for forgery detection.

4) B-Free (CVPR 2025) Guillaro et al. (2025). B-Free proposes a bias-free training paradigm by generating fake images from real images through image-conditioned diffusion models. This construction aligns the semantic content of real and fake samples, encouraging the detector to focus on generation artifacts rather than semantic or dataset biases. It further improves robustness through content-based augmentation.

5) DDA (NeurIPS 2025) Chen et al. (2025). DDA demonstrates that pixel-domain alignment alone is insufficient for fully aligning real and synthetic image pairs. Thus, they propose to align synthetic images with real ones across both pixel and frequency domains, thereby mitigating bias.

6) MIRROR (arXiv 2026) Liu et al. (2026). MIRROR reformulates AI-generated image detection as a reference-comparison problem. It projects an input image onto a manifold-consistent real-image

Table 10: **Data Scale of AI-Generated Document Images.**

Evaluation Split	Method	Type	Time	Resolution	Data Scale
Text2Image	Z-Image-Turbo	Open	2025.12	$1024 \sim 1024$	1,000
	HiDream-O1-Image	Open	2026.05	$2048 \sim 2048$	1,000
	Seedream 4.5	Closed	2025.12	$1184 \sim 3616$	1,000
	Nano-Banana-2	Closed	2026.02	$338 \sim 2816$	1,000
	GPT-Image-2	Closed	2026.04	$724 \sim 2172$	1,000
Image-Edit	FLUX.1-Kontext	Open	2025.06	$512 \sim 1984$	2,000
	FLUX.2-Klein	Open	2026.04	$336 \sim 1616$	2,000
	Hunyuan-3-Instruct	Open	2026.01	$527 \sim 1992$	2,000
	LongCat-Image-Edit	Open	2025.12	$528 \sim 2000$	2,000
	OmniGen2	Open	2025.06	$336 \sim 3264$	2,000
	Qwen-Image-Edit	Open	2025.08	$336 \sim 3264$	2,000
	Step1X-Edit	Open	2025.11	$336 \sim 3264$	2,000
	Seedream 4.5	Closed	2025.12	$1088 \sim 4096$	1,992
	Seedream 5.0-Lite	Closed	2026.07	$1088 \sim 4096$	1,993
	Nano-Banana-2	Closed	2026.02	$271 \sim 1024$	2,000
Digital-Image-Edit	Qwen-Image-Edit	Open	2025.08	$1152 \sim 2464$	1,000
	Nano-Banana-2	Closed	2026.02	$608 \sim 1024$	1,000

reference through sparse linear combination and uses the resulting residuals as detection evidence, aiming to reduce reliance on generator-specific forgery cues.

7) OmniAID (ICML 2026) Guo et al. (2025). OmniAID aims to decouple semantic content from generation artifacts for universal AIGI detection. It adopts a mixture-of-experts framework with specialized semantic experts and a universal artifact expert, enabling the detector to model content-dependent flaws while preserving content-agnostic artifact representations.

8) PROBE (ICML 2026) Cao et al. (2026). PROBE improves detector generalization by actively exploring difficult regions of the generative space. Instead of treating the generator as a fixed data source, it uses the detector as a critic to guide manifold-level modifications and synthesize hard samples, which are then used to expose failure cases and refine the detector.

9) SICA (ICML 2026) Du et al. (2026). SICA unifies Deepfake, AI-generated image, image-manipulation, and document-forgery detection within a single detector. To mitigate artifact feature collapse caused by heterogeneous forensic domains, it uses a frozen CLIP backbone as a semantic reference and applies constrained low-rank adaptation to construct a unified yet discriminative artifact feature space.

A.5 AUC Results on AIGDoc

To further assess the discriminative ability of detectors on document images, we additionally report AUC as a threshold-independent evaluation metric. As shown in Table 9, existing detectors exhibit substantial performance variation across different document generation settings. While some methods achieve strong AUC on relatively distinguishable generators, their average performance remains limited, particularly for born-digital image editing.

PROBE achieves the highest average AUC of 76.7% among the original detectors. Moreover, the text-density-guided variant, PE-LoRA†, further improves the average AUC to 84.2%, yielding a gain of 9.4% compared with PE-LoRA. This result is consistent with the accuracy analysis in the main paper and further supports our finding that text-dense regions provide stronger discriminative evidence for AI-generated document image detection. In contrast, the low AUC values observed for many methods on Digital-Qwen-Image-Edit and Digital-Nano-Banana-2 confirm that born-digital document editing remains a particularly challenging scenario.

Overall Description:
Describe the whole image in detail, including the image type, main content, visual style, background, colors, objects, and any important visual details.

Layout Structure:
Describe the layout and document structure in detail. Include the positions of titles, paragraphs, tables, images, logos, stamps, signatures, headers, footers, columns, rows, margins, borders, and any repeated sections. Mention the approximate reading order from top to bottom and left to right.

Image Text:
Extract all readable text from the image as completely and accurately as possible. The extracted text must remain in the original language shown in the image. For example, Chinese text should be output in Chinese, English text should be output in English, Japanese text should be output in Japanese, and mixed-language text should preserve each language exactly as it appears. Preserve the original reading order, line breaks, punctuation, numbers, capitalization, spacing, and symbols. If the image contains tables or forms, keep the table/form structure in text form as much as possible. Do not translate, summarize, paraphrase, or rewrite the text. If a word or character is uncertain, write [unclear]. If no text is visible, write "No visible text."

Figure 4: **Prompt Templates for Image-to-Text Generation.**

A.6 Details of Generation Pipeline

Due to the high text density and complex layout structures of document images, simple category-level prompts are insufficient for producing realistic document-like images. We therefore use the large vision-language model InternVL3.5-38B Wang et al. (2025b) to generate fine-grained prompts from real document images, so that the synthesized images can better preserve document content, layout organization, and realistic forgery scenarios. Due to sensitive-content restrictions, Seedream-4.5 produces only 1,992 edited images.

Statistical Distribution of the Test Set. We construct three types of AI-generated document images: text-to-image generation, camera-captured document editing, and born-digital document editing. The generation process involves 13 open-source and closed-source generation settings, producing a total of 27k AI-generated images. Detailed statistics for each generation type and model are shown in Table 10.

Text-to-image Generation. We randomly select 600 real document images and use InternVL3.5-38B to generate detailed text prompts for each image. Each prompt follows a three-part structure: *Overall Description*, which characterizes the document type and visual appearance; *Layout Structure*, which specifies the arrangement of text blocks, tables, stamps, and other document elements; and *Image Text*, which records representative visible textual content. To improve prompt diversity while mitigating privacy leakage, we further use InternVL3.5-38B to rewrite and enrich the extracted text, replacing sensitive fields with plausible synthetic information. This process expands the prompt set from 600 to 1,000 entries. The prompt template and representative generated examples are shown in Fig. 4 and Fig. 6, respectively.

Image Editing. We use 2K real document images as source images for editing. To better approximate real-world forgery scenarios, we employ InternVL3.5-38B to identify text regions and locate sensitive or editable fields, such as amounts, dates, numbers, and key terms. Based on these recognized fields, we generate localized editing prompts for image editing models. The prompt template used for this process is shown in Fig. 5.

You are preparing text-localized image editing instructions for a document/photo image.

Analyze the input image carefully. Read all visible text and choose exactly one important text span that would be meaningful to edit. Prefer text related to money, totals, prices, dates, times, IDs, report numbers, invoice numbers, ticket numbers, order numbers, seat numbers, quantities, phone numbers, addresses, names of institutions, or other key document fields.

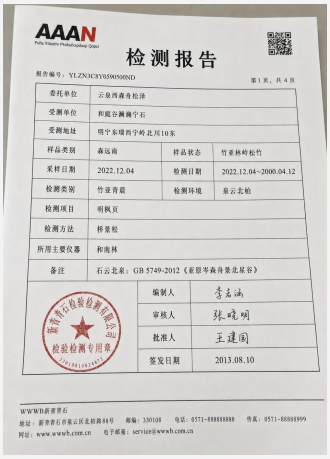
If several candidates exist, prioritize:
- amount/total/price/tax/payment values.
- date/time values.
- report/invoice/ticket/order/serial numbers.
- seat/flight/table/customer IDs.
- short title or institution name.

Create a realistic replacement for the selected span. Keep the replacement in the same language/script and format as the original selected text. Preserve currency symbols, date formats, punctuation, capitalization, and general length when possible. Do not select private sensitive personal data such as full ID-card numbers, bank card numbers, or phone numbers unless no other meaningful field is visible.

Respond with exactly one JSON object and nothing else:
- selected_text: <the exact text span to change, in the original language>,
- replacement_text: <the new text, same language/script/format>,
- recognized_context: <the surrounding recognized text line or short phrase that contains selected_text, in the original language>,
- edit_prompt: Change only the key text "<selected_text>" to "<replacement_text>".

Important rules:
- The JSON string values must be valid JSON strings.
- The edit_prompt itself must be in English, but selected_text, replacement_text, and recognized_context must remain in the original language shown in the image.
- Choose a compact span, not a whole paragraph. For example, choose "$24.20", "2018.05.10", "MU 9739", "Row 7 Seat 9", or "Report No.: BMBS1T2H5745755SZ".
- If no readable text is visible, choose a visually prominent short label and set recognized_context to "No visible text.".

Figure 5: **Prompt Templates for Image Editing.**



Generation model: GPT-Image-2
Prompt:{
"image_name":"0.jpg",
"overall_description":"The image shows a photographed Chinese testing report page from Prid Gabtawv Fgavvejvldllf Ytmcr. Vps paper is white with a printed table, a black and red MMMA logo at the upper left, a QR code at the upper right, and a red circular company seal near the lower left. The document is slightly tilted and photographed under uneven lighting. It appears to be the first page of a four-page drinking water test report.",
"layout_structure":"The page is vertically oriented. The top area contains the OOOI logo on the left, the centered title \"检测报告\", the report number on the upper left below the logo area, and the page count on the upper right. The main body is a large bordered table with rows for client, tested organization, address, sample category, sample status, sampling date, testing date, testing category, environment, testing items, methods, instruments, notes, reviewer fields, and issue date.",
"image_text":"AAAN\nPehs Xttodmr Pkxbufxspdasp Qopcr\n 检 测 报 告 \ n\n 报 告 编 号 ： YLZN3C8Y0590500ND\n第1页，共4页\n\n委托单位　云泉西森舟松泽\n受测单位　和庭谷澜澜宁石\n受测地址　明宁东瑞西宁岭北川10东\n样品类别　森远南　样品状态　竹亚林岭松竹\n采样日期　2022.12.04　检测日期　2022.12.04~2000.04.12\n检测类别　竹亚青晨检测环境　泉云北柏\n检测项目　明枫页\n检测方法　桥景松\n所用主要仪器　和南林\n备注　石云北泉：GB 5749-2012《亚原岑森舟景北星谷》\n\n编制人\n审核人\n批准人\n签发日期　2013.08.10\n\n\nWWWB新青青石\n[seal text partially unclear]"}



Generation model: Nano-Banana-2
Prompt:
{**"image_name"**:"157.jpg",
"overall_description":"The image shows a restaurant receipt from Cucx Kcdvh in Dtort Divv, California, printed on white paper on a wooden table. It lists order number, server, table/bar, date/time, three food items, subtotal, tax, total, order balance due, and suggested gratuity amounts. The receipt is clear and centered.",
"layout_structure":"The receipt is vertical. The top contains restaurant name, address, phone, and website. The upper right contains order number, server, table, and date/time. The middle item section lists items and prices. The bottom section contains subtotal, tax, total, order balance due, and a suggested gratuity table with percentage, tip, and total columns.",
"image_text":"Zxfg Flnny\n3313 Qnosmt Qlo. Gjk C21\nFvazu Dpjr, CA 21262\n644 693 2133\nwji.jfossivux.veq\n\nOrder# 222242\nServer: Jesa R\nTable: Bar 7\nDate: 1/23/18, 12:16 PM\n\nChicken $15.00\nPescado $18.00\nArrachera solo$9.50\n\nSubtotal: $42.50\nTotal Tax: $3.29\n\nTotal: $45.79\n\nRxgsa Jsoqbff due: $45.79\n\nIltotjwqq Squczjqa\n% Ahf Hmoum\n18.00% of sale: $7.65 = $53.44\n20.00% of sale: $8.50 = $54.29\n22.00% of sale: $9.35 =$55.14"}



Generation model: Seedream4.5
Prompt:
{**"image_name"**:"29.jpg",
"overall_description":"The image shows a Fgqhgbq Xcjhcjq Jmqvsvv Jmvqms Vvfde Rvnf ticket photographed close-up. The ticket has a red Omlhx Soavlgh Tibstsw logo, several lottery number rows, issue and sale information, address, QR code, prize pool line, previous winning numbers, and lower dot-matrix code area partly cut off by a green bar.",
"layout_structure":"The ticket is vertical but cropped at the bottom. The lottery logo and bilingual name are at the top. Game type, terminal/serial number, and a long code line appear below. The center lists five entries labeled A through A. Wap lower section contains issue date, price, sales time, address, QR code on the right, prize pool and previous winning number lines, and a partially hidden dot-matrix code.",
"image_text":"晨 景 西 东 瑞 西 \ nKXQHT VPXJQVH XGERARF\n\n 锦 河 : 景 溪 东 - 桥 景 00442241\n2C59-A23B-7B9C-EB1A-17CC/52401115/40BD2\nA.08 14 16 21 23 32-06 (1)\nB.07 13 15 18 21 25-07 (1)\nC.08 14 20 23 30 33-08 (1)\nD.07 14 20 23 24 25-09 (1)\nE.02 14 19 21 23 27-10 (1)\n\n泽青期:4441211 19-12-03　¥:10\n岑晨期:4441211-243　19-12-03 13:20:12\n地址:星竹云桥区和亚景溪桥锦和A116\n舟禾亚明:475055125.00元,138\n期庭星号晨:06-32-31-30-02-12\n[QR code]\n[dot-matrix code cropped]"}

Figure 6: **Prompt Examples and Generated Images.**