Retrieval Augmented Visual Detection: A Knowledge-Driven Approach for AI-Generated Image Identification

Mamadou Keita[1]
https://www.uphf.fr/
Wassim Hamidouche[2]
https://www.ku.ac.ae/
Hessen Bougueffa Eutamene[1]
https://www.uphf.fr/
Abdelmalik Taleb-Ahmed[1]
https://www.uphf.fr/
Abdenour Hadid[3]
https://www.sorbonne.ae

[1] Laboratory of IEMN, Polytechnic University of Hauts-de-France Valenciennes, France
[2] Khalifa University Abu Dhabi, UAE
[3] Sorbonne Center for Artificial Intelligence Sorbonne University Abu Dhabi, UAE

Abstract

We introduce the first framework for AI-generated image detection that leverages visual retrieval-augmented generation (RAG). We propose to dynamically retrieve relevant images to enhance detection by utilizing a fine-tuned CLIP image encoder, enhanced with category-related prompts to improve representation learning. We further integrate a vision-language model (VLM) to fuse retrieved images with the query, enriching the input and improving accuracy. Given a query image, our proposed approach generates an embedding, retrieves the most relevant images from a database, and combines these with the query image to form an enriched input for a VLM. Experiments on the UniversalFakeDetect benchmark, which covers 19 generative models, show that our approach achieves state-of-the-art performance with an average accuracy of 93.85%. It outperforms traditional methods in terms of robustness, maintaining high accuracy even under image degradations such as Gaussian blur and JPEG compression. Specifically, it achieves an average accuracy of 80.27% under degradation conditions, compared to 63.44% for the state-of-the-art model C2P-CLIP, demonstrating consistent improvements in both Gaussian blur and JPEG compression scenarios. Our approach also shows strong cross-domain generalization, achieving 78.81% average accuracy on diverse unseen data, confirming its effectiveness in open-world scenarios.

1 Introduction

The rapid advancement of generative models, particularly in image synthesis, has introduced significant challenges in distinguishing AI-generated content from real data. For instance, generative adversarial networks (GANs) [6, 9, 10] and diffusion-based models [8, 19, 22,

23] have become increasingly proficient at producing photorealistic images that are nearly indistinguishable from genuine ones. However, the progress in detection methods has not kept pace with these advancements, creating an urgent need for robust and reliable detectors

Traditional AI-generated image detection approaches primarily rely on identifying low-level artifacts or model-specific fingerprints [25, 26, 32] present in synthetic images. These artifacts include pixel inconsistencies, noise patterns, and subtle distortions that reveal traces of the underlying generation process. While these methods have demonstrated effectiveness in controlled settings, they often fail in real-world scenarios. As generative models improve, they become more adept at minimizing artifacts and replicating the statistical properties of real images, making detection increasingly difficult. Furthermore, many existing detection methods suffer from a fundamental limitation: over-reliance on model-specific features and low-level artifacts. Since these methods are often tailored to exploit weaknesses in particular architectures, they struggle to generalize across different generative models. Consequently, there is a growing need for more adaptive detection approaches that leverage additional sources of information to enhance the performance, robustness, and reliability of the detectors. One appealing direction is Retrieval-Augmented Generation (RAG) [14], a paradigm initially developed to improve factual accuracy in large language models by retrieving and incorporating external knowledge relevant to a given query. While extensively explored in textual tasks, its potential for visual tasks, particularly AI-generated image detection, remains largely underexplored.

Hence, we propose an approach called RAVID as a novel retrieval-augmented framework for AI-generated image detection. Unlike traditional methods that rely solely on model-dependent features, RAVID retrieves visually similar images relevant to the input query and integrates them into the detection process, thereby enhancing accuracy and robustness. At its core, RAVID leverages a Contrastive Language-Image Pretraining (CLIP)-based image encoder, fine-tuned through category-level prompt integration to improve its ability to capture semantic and structural patterns crucial for distinguishing AI-generated images from real ones. Additionally, we incorporate vision-language models (VLMs), such as Open-flamingo [1], to effectively fuse retrieved images with the query input, enabling richer contextual understanding. By combining retrieval-based augmentation with advanced vision-language decision-making ability, our approach significantly improves adaptability and effectiveness, making it well-suited for real-world applications where reliable AI-generated content detection is critical.

To evaluate our approach, we conducted extensive experiments on UniversalFakeDetect, a large-scale benchmark comprising AI-generated images from 19 generative models. Experimental results demonstrated that RAVID consistently outperforms existing detection methods, achieving an average accuracy of 93.85% and exhibiting better generalization across multiple challenging settings. In addition, RAVID demonstrated strong resilience to image degradations, such as Gaussian blur and JPEG compression, significantly outperforming the state-of-the-art C2P-CLIP model [27]. Unlike traditional methods, our approach maintains high accuracy by leveraging retrieval-augmented generation to compensate for lost visual features, ensuring robust performance even under real-world distortions. This highlights the effectiveness of retrieval-augmented techniques in enhancing generalization and robustness across different image generators.

Our key contributions are summarized in: (**i**) A novel retrieval-augmented framework for AI-generated image detection, dynamically retrieving and integrating external visual knowledge to enhance decision-making; (**ii**) Fine-tuning of a CLIP-based image encoder with category-level prompt integration, improving representation learning for retrieval tasks;

and (**iii**) Integration of VLMs to combine retrieved images with queries, enhancing contextual understanding and detection robustness.

2 Proposed Method: RAVID

In this section, we present RAVID, which performs retrieval of query-relevant images over vector image corpus as the external knowledge source and determines query-image class (label) grounded in them, as illustrated in Figure 1.

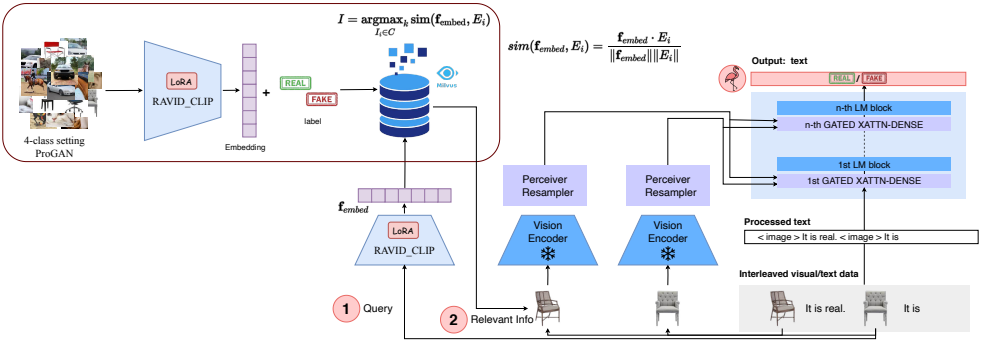

Figure 1: RAVID integrates a fine-tuned CLIP-based image encoder (RAVID_CLIP) for embedding-based image retrieval and Openflamingo for decision-making: (a) 4-class ProGAN training set images are encoded into vector embeddings using RAVID_CLIP and stored in a Milvus vector database; (b) At testing time, the query image embedding is matched against stored embeddings to retrieve the most relevant images; (c) The retrieved images and labels serve as contextual information, combined with the query image, and processed by Openflamingo.

2.1 Concept-Aware Image Embeddings

Recent work by Tan *et al.* [27] demonstrates that CLIP features' ability to detect AI-generated images through linear classification is largely due to their capacity to capture *conceptual similarities*. Building on this insight, they propose an approach that incorporates enhanced captions and contrastive learning to embed categorical concepts into the image encoder, thereby improving the distinction between real and generated images. Inspired by this work, we adopt a similar strategy in our framework to generate high-quality embeddings for a vector database.

Caption Generation and Enhancement. Let $\mathcal{D}$ represent the training dataset containing both real and synthetic images, defined as $\mathcal{D} = \{(x_j, y_j)\}_{j=1}^{N}$, where $y_j \in \{0,1\}$ indicates whether the image x_j is real ($y=0$) or fake ($y=1$). For each image in the dataset, we generate captions using the ClipCap model [17], resulting in a set of captions $\mathcal{C} = \{(c_j, y_j)\}_{j=1}^{N}$.

To enhance these captions, we append category-specific prompts $\mathcal{P} = \{P_{\text{real}}, P_{\text{fake}}\}$ to the original captions, as proposed by [27]. Specifically, the enhanced captions $\tilde{\mathcal{C}} = \{\tilde{c}_j\}_{j=1}^{N}$ are defined as:

$$\tilde{c}_j = \begin{cases} (P_{\text{real}}, c_j), & \text{if } y_j = 0 \\ (P_{\text{fake}}, c_j), & \text{if } y_j = 1 \end{cases} \tag{1}$$

In this formulation, P_{real} and P_{fake} represent category-specific prompts (e.g., $P_{\text{real}} = \text{Camera}$, $P_{\text{fake}} = \text{Deepfake}$) that provide additional context to differentiate real images from synthetic ones.

Concept Injection via Contrastive Learning. To integrate classification concepts into the image encoder, we employ a contrastive learning framework. In this approach, the text encoder remains frozen, while Low-Rank Adaptation (LoRA) layers are applied to the image encoder to facilitate learning. Given an image x_j and its corresponding enhanced caption $\tilde{c}_j$, their feature representations are computed as follows:

$$t_j = \text{encoder}_{\text{text}}(\tilde{c}_j), \quad \mathbf{e}_j = \text{encoder}_{\text{img}}(x_j), \tag{2}$$

where t_j and $\mathbf{e}_j$ denote the text and image embeddings, respectively.

To ensure that the image encoder aligns visual features with their corresponding textual descriptions, we optimize a contrastive loss function $L_{\text{contrastive}}$, defined as:

$$L_{\text{contrastive}} = \frac{1}{2}\left(L_{\mathbf{e}\to t} + L_{t\to\mathbf{e}}\right), \tag{3}$$

where $L_{\mathbf{e}\to t}$ enforces alignment between image embeddings and their respective text embeddings, while $L_{t\to\mathbf{e}}$ ensures reverse alignment. These losses are formulated as:

$$L_{\mathbf{e}\to t} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{e}_i^T \cdot t_i)}{\sum_{j=1}^{N}\exp(\mathbf{e}_i^T \cdot t_j)}, \tag{4}$$

$$L_{t\to\mathbf{e}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(t_i^T \cdot \mathbf{e}_i)}{\sum_{j=1}^{N}\exp(t_i^T \cdot \mathbf{e}_j)}. \tag{5}$$

Here, $\mathbf{e}_i^T t_j$ represents the dot product between the image feature $\mathbf{e}_i$ and the text feature t_j, capturing their similarity. The denominator normalizes the similarity scores across all samples in the batch, ensuring a well-structured representation space.

By optimizing $L_{\text{contrastive}}$, the image encoder learns to map visual features into a space that aligns with their textual descriptions. This process effectively injects classification concepts within the image encoder, enhancing its ability to distinguish between real and AI-generated images.

The CLIP concept-injection-enhanced model is then used to generate embeddings for both real and fake images. These embeddings are stored in a vector database (e.g., Milvus [30]), which serves as an external knowledge source for the RAVID framework.

2.2 Image Retrieval

Retrieval involves computing the similarity between the query image q_{img} and each knowledge element to determine relevance. To achieve this, we first embed the query image q_{img} using the RAVID_CLIP image encoder to obtain its embedding $\mathbf{f}_{\text{embed}}$. Relevance is then computed based on representation-level similarity, such as cosine similarity, to measure the alignment between the query embedding and stored embeddings in the external corpus C. The retrieval process is formulated as:

$$I = \underset{I_i \in C}{\text{argmax}_k}\, \text{sim}(\mathbf{f}_{\text{embed}}, E_i), \tag{6}$$

where $\mathbf{f}_{\text{embed}}$ is the embedding of the query image q_{img} computed by the RAVID_CLIP image encoder, E_i represents the stored embedding of an image I_i in the external corpus C, and $\text{sim}(\mathbf{f}_{\text{embed}}, E_i)$ denotes the cosine similarity between the query embedding and each corpus embedding, computed as:

$$\text{sim}(\mathbf{f}_{\text{embed}}, E_i) = \frac{\mathbf{f}_{\text{embed}} \cdot E_i}{\|\mathbf{f}_{\text{embed}}\| \|E_i\|} \tag{7}$$

where $\underset{I_i \in C}{\text{argmax}_k}$ selects the top-k images with the highest similarity scores.

By retrieving the top-k most relevant images, this approach ensures that the subsequent answer generation step benefits from rich contextual information, improving the robustness of AI-generated image detection.

2.3 Image-Augmented Response Generation

After retrieving the most relevant images, the next step is to integrate them into the response generation process to enhance the quality and contextual accuracy of the generated output. To achieve this, we first construct a multimodal context by pairing each retrieved image with its corresponding label. These multimodal pairs are then concatenated to form a comprehensive context representation. Finally, the query image is incorporated into this structured input, which serves as the input to a VLM, such as Openflamingo. Formally, this process is represented as:

```
[
    {'text': 'Is this photo real? Please provide your answer. You should
        ONLY output "real" or "fake".'},
    {'image': 'path to img_1'},
    {'text': 'User: It is \nAssistant: img_1_label'},
    {'image': 'path to img_2'},
    {'text': 'User: It is \nAssistant: img_2_label'},
    ...
    {'image': 'path to img_n'},
    {'text': 'User: It is \nAssistant: img_n_label'},
    {'image': 'path to q_img'},
    {'text': 'User: It is \nAssistant: '}
]
```

This structured input is then fed into the VLM, which jointly processes visual, textual, and query-specific information. By leveraging this multimodal richness, the model generates a response that effectively integrates retrieved knowledge to improve AI-generated image detection accuracy.

3 Experimental Analysis

Dataset. To ensure a fair comparison, we utilize the widely adopted UniversalFakeDetect dataset [20], extensively used in prior benchmarks, allowing for a direct evaluation of RAVID against state-of-the-art methods, ensuring consistency and robustness in performance assessment. Following the experimental setup introduced by Wang *et al.* [31], the dataset uses ProGAN as a training set, comprising 20 subsets of generated images. To build our vector database and fine-tuning CLIP ViT-L/14, we use a 4-class setting (horse, chair, cat, car) as indicated by Tan *et al.* [27]. The test set consists of 19 subsets generated by a diverse range

of models, including ProGAN [9], StyleGAN [10], BigGAN [10], CycleGAN [10], StarGAN [6], GauGAN [10], Deepfake [24], CRN [5], IMLE [15], SAN [7], SITD [4], Guided Diffusion [8], LDM [23], GLIDE [19], and DALLE [22].

To further evaluate RAVID's generalization capability, we compare it against the best-performing models trained on the ProGAN 4-class setup using a different testing dataset. This dataset includes 12 subsets: 2 real data subsets (MS COCO and Flickr) and 10 synthetic subsets (ControlNet, DALL·E 3, DiffusionDB, IF, LaMA, LTE, SD2Inpaint, SDXL, SGXL, and SD3).

Evaluation Metrics. Following the convention of previous detection methods [11, 20, 27, 32], we report the accuracy (ACC). We also calculate the mean accuracy across all data subsets to provide a more comprehensive evaluation of overall model performance.

Baselines. In our study, we fine-tuned AntiFakePrompt [3] and Bi-LORA [12]. Moreover, we have chosen the latest and most competitive methods in the field, including Co-occurence [18], Freq-spec [33], CNN-Spot [31], FatchFor [2], UniFD [20], LGrad [26], F3Net [21], FreqNet [28], NPR [29], Fatformer [16], C2P-CLIP [27], RINE [13]. For all these models, we report the results presented in [27]. For RINE, we report the results from its paper [13].

Implementation Details. To fine-tune CLIP ViT-L/14, we use Adam optimizer with an initial learning rate of 4×10^{-4}, a batch size of 128, and train for one epoch. We apply LoRA layers to the q_proj, k_proj, and v_proj layers using the Parameter-Efficient Fine-Tuning (PEFT) library. The LoRA hyper-parameters are as follows: $lora\_r = 6$, $lora\_alpha = 6$, and $lora\_dropout = 0.8$. For the vector database, we use Milvus locally via Docker. On the other hand, for image-augmented response generation, we use Openflamingo [1].

Comparison with the State-of-the-Art. Table 1 presents the mean accuracy (mAcc) scores for cross-generator detection on the UniversalFakeDetect dataset, which includes 19 different generative models spanning GANs, Deepfakes, low-level vision models, perceptual loss models, and diffusion models. RAVID achieves 93.85% mAcc, outperforming 15 state-of-the-art methods and demonstrating strong generalization across diverse image synthesis techniques.

Compared to UniFD, a recent state-of-the-art method, RAVID improves the mAcc by 12.47%, highlighting the effectiveness of our approach. Additionally, RAVID demonstrates competitive performance with the latest methods, RINE and C2P-CLIP, achieving a mere 1.48% and 0.16% mAcc gap, respectively. While RINE utilizes advanced feature extraction and fusion techniques, which increase computational complexity, C2P-CLIP embeds category-specific concepts into CLIP's image encoder. Meanwhile, our method strikes a balance between performance and efficiency, making it more suitable for real-world applications.

Impact of Retrieved Image Count on Detection Performance. To evaluate the impact of the number of retrieved images on RAVID's AI-generated image detection performance, we conducted experiments with varying retrieval settings. Specifically, we compared performance when retrieving 1 ($N = 1$), 3 ($N = 3$), 5 ($N = 5$), 7 ($N = 7$), and 13 ($N = 13$) images, utilizing the Openflamingo vision-language model. The results, presented in Table 2, show a

Table 1: Accuracy (ACC) scores of state-of-art detectors and RAVID across 19 test datasets. Best performance is denoted with **bold**. We report the results of the best RAVID models with different VLMs.

Methods	Ref	GAN						Deep Fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	mAcc
		Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/cfg	100 steps	100 27	50 27	100 10		
Freq-spec	WIFS2019	49.90	**99.90**	50.50	49.90	50.30	99.70	50.10	50.00	48.00	50.60	50.10	50.90	50.40	50.40	50.30	51.70	51.40	50.40	50.00	55.45
Co-occurence	Elect. Imag.	97.70	97.70	53.75	92.50	51.10	54.70	57.10	63.06	55.85	65.65	65.80	60.50	70.70	70.55	71.00	70.25	69.60	69.90	67.55	66.86
CNN-Spot	CVPR2020	99.99	85.20	70.20	85.70	78.95	91.70	53.47	66.67	48.69	86.31	86.26	60.07	54.03	54.96	54.14	60.78	63.80	65.66	55.58	69.58
Patchfor	ECCV2020	75.03	68.97	68.47	79.16	64.23	63.94	75.54	75.14	75.28	72.33	55.30	67.41	76.50	76.10	75.77	74.81	73.28	68.52	67.91	71.24
F3Net	ECCV2020	99.38	76.38	65.33	92.56	58.10	**100.00**	63.48	54.17	47.26	51.47	51.47	69.20	68.15	75.35	68.80	81.65	83.25	83.05	66.30	71.33
Bi-LORA	ICASSP2023	98.71	96.74	81.18	78.30	96.30	86.32	57.78	68.89	52.28	73.00	82.60	65.10	85.15	59.20	85.00	83.50	85.65	84.90	72.70	78.59
LGrad	CVPR2023	99.84	85.39	82.88	94.83	72.45	99.62	58.00	62.50	50.00	50.74	50.78	77.50	94.20	95.85	94.80	87.40	90.70	89.55	88.35	80.28
UniFD	CVPR2023	**100.00**	98.50	94.50	82.00	99.50	97.00	66.60	63.00	57.50	59.50	72.00	70.03	94.19	73.76	94.36	79.07	79.85	78.14	86.78	81.38
AntiFakePrompt	CVPR2023	99.26	96.82	87.88	80.00	98.13	83.57	60.20	70.56	53.70	79.21	79.01	73.75	89.55	64.10	89.80	93.55	93.90	92.95	80.10	82.42
FreqNet	AAAI2024	97.90	95.84	90.45	**97.55**	90.24	93.41	**97.40**	88.92	59.04	71.92	67.35	**86.70**	84.55	**99.58**	65.56	85.69	**97.40**	88.15	59.06	85.09
NPR	CVPR2024	99.84	95.00	87.55	96.23	86.57	99.75	76.89	66.94	**98.63**	50.00	50.00	84.55	97.65	98.00	98.20	**96.25**	97.15	**97.35**	87.15	87.56
FatFormer	CVPR2024	99.89	99.32	**99.50**	97.15	99.41	99.75	93.23	81.11	68.04	69.45	69.45	76.00	98.60	94.90	98.65	94.35	94.65	94.20	**98.75**	90.86
RINE	ECCV2024	100.00	99.30	99.60	88.90	**99.80**	99.50	80.60	90.60	68.30	89.20	90.60	76.10	98.30	88.20	98.60	88.90	92.60	90.70	95.00	91.31
C2P-CLIP*	AAAI2025	99.71	90.69	95.28	**99.38**	95.26	96.60	89.86	**98.33**	64.61	90.69	90.69	77.80	99.05	98.05	98.95	94.65	94.20	94.40	**98.80**	93.00
C2P-CLIP‡	AAAI2025	99.98	97.31	99.12	96.44	99.17	99.60	**93.77**	95.56	64.38	**93.29**	93.29	69.10	**99.25**	97.25	99.30	95.25	95.25	96.10	98.55	93.79
RAVID (N=13)	-	99.98	97.35	99.15	96.27	99.33	99.82	93.47	95.00	63.70	95.53	95.56	68.75	99.20	96.95	**99.35**	94.65	94.90	95.85	98.35	**93.85**

(⋆) Trump,Biden. (‡) Deepfake,Camera.

substantial improvement in detection accuracy as the number of retrieved images increases. These findings indicate that incorporating additional retrieved images provides richer contextual information, thereby improving the model's generalization across diverse AI-generated image models.

Table 2: Impact of Retrieved Image Count on Detection Performance.

Methods	VLMs	N Shots	GAN						Deep Fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	mAcc
			Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/cfg	100 steps	100 27	50 27	100 10		
RAVID W/ RAG	Openflamingo	1	50.00	50.00	50.00	50.00	50.00	50.00	50.08	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00
RAVID W/ RAG	Openflamingo	3	99.95	97.84	99.25	95.94	99.38	99.85	92.64	93.61	62.33	97.55	97.59	66.95	99.35	96.35	99.35	93.10	93.25	94.40	98.40	93.53
RAVID W/ RAG	Openflamingo	5	99.95	97.58	99.28	96.24	99.30	99.82	93.19	93.89	63.24	96.26	96.25	68.10	99.20	96.45	99.30	94.00	94.35	95.10	98.30	93.67
RAVID W/ RAG	Openflamingo	7	99.96	97.46	99.15	96.24	99.32	99.80	93.30	94.72	63.93	96.04	96.03	68.30	99.25	96.65	99.35	94.00	94.70	95.30	98.15	93.77
RAVID W/ RAG	Openflamingo	13	99.98	97.35	99.15	96.27	99.33	99.82	93.47	95.00	63.70	95.53	95.56	68.75	99.20	96.95	99.35	94.65	94.90	95.85	98.35	93.85

Impact of Using RAG for Retrieval in RAVID. To evaluate the impact of dynamically retrieving relevant images in the RAVID approach, we conducted an experiment where the context provided to the VLM was formed by randomly selecting images, instead of using the RAG retrieval mechanism. This experiment allowed us to assess the significance of the retrieval process in improving detection accuracy. In this setup, rather than retrieving relevant images related to the query image, we randomly selected N images from the 4-class setting ProGAN training set. These randomly chosen images, along with their corresponding labels, were used as a context for the detection task. This mimicked the in-context learning strategy used in RAG, but without its retrieval component. We maintained the same configurations as in RAVID, varying the number of selected images (shots). In the 3-shot setup, three randomly selected image was provided as context, while in the 13-shot setup, thirteen images were used as context.

The results in Table 3 show the detection accuracy (*mAcc*) across a range of generative models. For the 3-shot setup, RAVID W/O RAG achieved an average accuracy of 50.10%, whereas in the 13-shot setup, the accuracy slightly declined to 49.90%. While these results indicate that providing more context does not improve performance, they still fall behind the accuracy achieved when using RAG, where the retrieved context is more relevant to the query image. This experiment highlights the importance of relevant context in AI-generated image detection. When the model relies on randomly selected images, the context lacks meaningful relevance to the query, limiting its ability to make accurate predictions, especially with complex generative models. In contrast, the ability of RAG to retrieve relevant images sub-

stantially boosts the model's performance, emphasizing the critical role of relevant context in improving detection accuracy and generalization. This investigation also underscores the need for an image embedding model for retrieval that is sensitive to the subtle characteristics of AI-generated images, as opposed to one that focuses on general cues, which are less useful for this task.

Table 3: In-context learning without RAG. Instead of retrieving relevant images to the query image, we randomly select N images from the 4-class setting ProGAN training set. Best performance is denoted with **bold**.

Methods	VLMs	N Shots	GAN						Deep Fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	mAcc
			Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/cfg	100 steps	100 27	50 27	100 10		
RAVID W/ RAG	Openflamingo	3	99.95	97.84	99.25	95.94	99.38	99.85	92.64	93.61	62.33	97.55	97.59	66.95	99.35	96.35	99.35	93.10	93.25	94.40	98.40	93.53
RAVID W/ RAG	Openflamingo	13	99.98	97.35	99.15	96.27	99.33	99.82	93.47	95.00	63.70	95.53	95.56	68.75	99.20	96.95	99.35	94.65	94.90	95.85	98.35	93.85
RAVID W/O RAG	Openflamingo	3	49.49	50.53	51.00	49.99	49.64	50.88	50.53	49.17	45.89	49.89	49.87	51.40	50.65	50.40	50.35	50.30	50.45	50.95	50.45	50.10
RAVID W/O RAG	Openflamingo	13	49.38	51.59	50.78	49.67	49.80	50.55	51.14	50.00	47.72	50.65	49.85	48.50	49.25	49.85	48.90	49.70	50.50	51.55	48.80	49.90

Robustness to Image Degradation.

To systematically evaluate the robustness of our proposed approach, we assess its performance under two common forms of image degradation: Gaussian blur and JPEG compression. These perturbations simulate real-world challenges where images undergo quality loss due to compression artifacts or motion blur, which can adversely impact AI-generated image detection. We compare our method, RAVID (N=13) W/ RAG Openflamingo, against the baseline C2P-CLIP, analyzing their degradation trends across multiple generative models. The quantitative results are summarized in Table 4, and the performance trends under different blur and JPEG compression levels are illustrated in Figure 2.

Table 4: Performance (ACC) after applying common image degradations.

Methods	VLMs	Degradation	GAN						Deep Fakes	Low level		Perceptual loss		Guided	LDM			Glide			Dalle	mAcc
			Pro-GAN	Cycle-GAN	Big-GAN	Style-GAN	Gau-GAN	Star-GAN		SITD	SAN	CRN	IMLE		200 steps	200 w/cfg	100 steps	100 27	50 27	100 10		
C2P-CLIP	-	Blur $\sigma = 1$	96.10	90.31	97.02	97.00	95.75	96.80	93.43	95.56	57.08	68.84	68.84	47.90	01.20	06.60	01.60	12.50	13.30	10.60	01.60	55.37
C2P-CLIP	-	Blur $\sigma = 2$	72.20	85.24	87.35	79.45	90.08	86.47	80.33	95.56	51.60	60.90	61.17	45.40	01.20	06.60	01.60	12.50	13.30	10.60	03.50	49.74
C2P-CLIP	-	Blur $\sigma = 3$	77.20	84.18	77.03	62.87	87.19	88.64	75.52	95.00	49.54	56.39	56.39	47.80	13.50	35.30	12.20	42.00	40.40	42.20	13.70	55.63
RAVID (N=13)	Openflamingo	Blur $\sigma = 1$	96.50	90.46	97.28	96.90	96.17	97.47	93.23	95.00	56.85	74.12	74.11	73.35	97.40	94.30	97.25	91.25	90.90	92.75	97.00	89.59
RAVID (N=13)	Openflamingo	Blur $\sigma = 2$	73.73	85.47	87.83	79.99	91.12	88.44	81.35	95.00	51.37	65.37	66.07	73.40	94.05	88.65	94.30	80.80	82.05	81.10	94.30	81.81
RAVID (N=13)	Openflamingo	Blur $\sigma = 3$	78.51	84.56	76.78	63.35	88.10	90.17	76.47	95.00	49.54	59.97	60.15	70.30	87.80	75.70	88.25	73.30	74.25	73.20	87.25	76.46
C2P-CLIP	-	Jpeg $q = 80$	95.80	94.93	92.92	75.60	96.85	95.02	85.57	94.72	55.94	92.53	92.42	64.80	14.30	53.10	16.80	58.40	63.40	56.80	17.10	69.32
C2P-CLIP	-	Jpeg $q = 70$	94.49	94.44	87.10	65.30	95.18	92.72	84.90	93.89	54.79	86.45	88.09	70.30	24.10	67.70	27.90	61.70	64.10	56.70	30.50	70.54
C2P-CLIP	-	Jpeg $q = 60$	94.59	94.40	81.80	62.30	95.57	89.62	82.76	90.56	53.65	80.00	80.05	65.90	23.20	69.70	26.10	58.00	60.50	54.50	42.10	68.70
C2P-CLIP	-	Jpeg $q = 50$	93.79	93.26	80.55	60.17	94.69	92.12	78.74	88.06	52.97	76.64	73.89	63.40	25.70	71.40	25.20	59.90	62.60	56.50	45.40	68.16
C2P-CLIP	-	Jpeg $q = 40$	93.23	91.79	75.55	57.20	93.22	92.62	75.10	81.39	52.74	77.96	75.56	71.50	35.10	77.80	34.60	64.90	65.10	62.20	52.80	70.02
RAVID (N=13)	Openflamingo	Jpeg $q = 80$	95.90	95.31	92.68	75.07	97.00	96.00	84.14	93.61	55.71	92.96	93.32	66.15	91.20	70.95	89.90	69.20	66.65	70.15	89.80	83.46
RAVID (N=13)	Openflamingo	Jpeg $q = 70$	94.24	94.55	86.58	65.05	95.09	93.85	84.16	93.33	54.57	83.85	86.38	63.65	86.20	64.60	84.25	68.00	66.50	70.20	82.55	79.87
RAVID (N=13)	Openflamingo	Jpeg $q = 60$	94.49	94.47	81.23	62.09	95.42	91.50	82.26	89.72	53.20	82.07	82.23	65.55	86.85	63.35	85.45	69.45	68.20	71.30	77.10	78.73
RAVID (N=13)	Openflamingo	Jpeg $q = 50$	93.83	93.30	80.05	60.07	94.58	93.72	77.91	87.22	52.28	78.78	76.04	66.80	85.55	62.10	85.70	68.75	67.60	70.15	74.90	77.33
RAVID (N=13)	Openflamingo	Jpeg $q = 40$	93.01	91.98	74.98	57.05	92.99	92.90	74.01	80.83	52.51	77.20	75.08	63.70	81.65	60.40	81.60	66.90	66.50	68.10	72.30	74.93
C2P-CLIP	-	Average	89.68	91.07	**84.91**	**69.99**	93.57	91.75	**82.04**	**91.84**	**53.54**	74.96	74.55	59.62	17.29	48.52	18.25	46.24	47.84	43.76	25.84	63.44
RAVID (N=13)	Openflamingo	Average	**90.03**	**91.26**	84.68	69.95	**93.81**	**93.01**	81.69	91.21	53.25	**76.79**	**76.67**	**67.86**	**88.84**	**72.51**	**88.34**	**73.46**	**72.83**	**74.62**	**84.4**	**80.27**

Across both degradation types, RAVID consistently demonstrates greater robustness compared to the baseline. This can be attributed to its retrieval-augmented mechanism, which enhances contextual understanding by leveraging external image priors. By integrating relevant external information, RAVID maintains high detection accuracy even when the image is degraded, as this additional information helps counter the impact of quality-reducing transformations without needing to recover lost details. Specifically, RAVID achieves an average accuracy of 80.27% under degradation conditions, compared to 63.44% for the state-of-the-art model C2P-CLIP, demonstrating consistent improvements in Gaussian blur and JPEG compression scenarios.

Generalization on Unseen Data. To assess cross-domain robustness, we evaluated the generalization performance of four top-performing methods from Table 1, including FatFormer, RINE, C2P-CLIP, and RAVID. Each model is trained solely on the ProGAN 4-class dataset

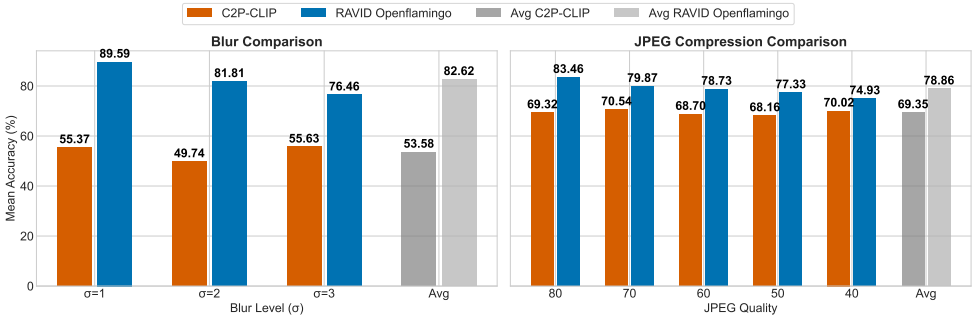

Figure 2: RAVID's robustness under gaussian blur and JPEG compression, common real-world degradations affecting AI-generated image detection.

and tested on a broad spectrum of unseen data sources, including authentic images (e.g., MS COCO, Flickr) and diverse generative models. This evaluation simulates a realistic open-world scenario where detection models face data distributions that deviate significantly from their training domain. As shown in Table 5, RAVID exhibits strong cross-domain generalization, achieving the highest overall mean accuracy of 78.81% across both real and synthetic domains.

Table 5: Generalization performance of methods trained on 4-class ProGAN. Results show accuracy (%) on real and synthetic data subsets, each containing 3,000 image samples.

Methods	#params	MS COCO	Flickr	ControlNet	Dalle3	DiffusionDB	IF	LaMA	LTE	SD2Inpaint	SDXL	SGXL	SD3	mAcc
FatFormer	493M	33.97	34.04	28.27	32.07	28.10	27.95	28.67	12.37	22.63	31.97	22.23	35.91	28.18
RINE	434M	**99.80**	**99.90**	**91.60**	75.00	**73.00**	77.40	30.90	98.20	71.90	22.20	98.50	08.30	70.56
C2P-CLIP	304M	99.67	99.73	15.10	**75.57**	27.87	**89.56**	**65.43**	00.20	27.90	**82.90**	07.17	**70.46**	55.13
RAVID (N=13)	-	97.83	99.23	85.80	68.93	70.70	60.71	62.97	99.97	80.37	62.10	98.80	58.31	78.81

4 Conclusion

We introduced RAVID, a novel retrieval-augmented framework for detecting AI-generated images. By dynamically retrieving and integrating relevant visual knowledge, RAVID enhances detection accuracy and generalization. Unlike traditional methods that rely on low-level artifacts or model fingerprints, our approach leverages a fine-tuned CLIP-based image encoder for embedding generation and retrieval. Additionally, we incorporate VLMs like Openflamingo to enrich contextual understanding. Evaluations on the UniversalFakeDetect benchmark showed that RAVID outperforms existing methods, achieving 93.85% accuracy in both in- and out-of-domain settings. A detailed analysis revealed a 35.51% performance gap between setups W and W/O retrieval in the 3-shot setting, highlighting the critical role of relevant context. Our findings confirm that retrieval not only enhances accuracy but also scales with additional retrieval shots, reinforcing its impact. Furthermore, our robustness analysis demonstrated that RAVID maintains superior detection performance under common image degradations, including Gaussian blur and JPEG compression. These results highlight the resilience of retrieval-based approaches in real-world conditions where images often suffer from quality loss due to compression pipelines, motion blur, or other distortions.

Acknowledgments: This work has been partially funded by the project PCI2022-134990-2 (MARTINI) of the CHISTERA IV Cofund 2021 program. Abdenour Hadid is funded by TotalEnergies collaboration agreement with Sorbonne University Abu Dhabi.

References

[1] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[2] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 103–120. Springer, 2020.

[3] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *CoRR*, 2023.

[4] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3291–3300, 2018.

[5] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017.

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[7] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11065–11074, 2019.

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

[9] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[11] Mamadou Keita, Wassim Hamidouche, Hassen Bougueffa, Abdenour Hadid, and Abdelmalik Taleb-Ahmed. Harnessing the power of large vision language models for synthetic image detection. *arXiv preprint arXiv:2404.02726*, 2024.

[12] Mamadou Keita, Wassim Hamidouche, Hessen Bougueffa Eutamene, Abdelmalik Taleb-Ahmed, David Camacho, and Abdenour Hadid. Bi-lora: A vision-language approach for synthetic image detection. *Expert Systems*, 42(2):e13829, 2025.

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

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[15] Ke Li, Tianhao Zhang, and Jitendra Malik. Diverse image synthesis from semantic layouts via conditional imle. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4220–4229, 2019.

[16] 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*, pages 10770–10780, 2024.

[17] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[18] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, Amit K Roy-Chowdhury, and BS Manjunath. Detecting gan generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*, 2019.

[19] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.

[20] 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*, pages 24480–24489, 2023.

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

[22] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[23] 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*, pages 10684–10695, 2022.

[24] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

[25] Sergey Sinitsa and Ohad Fried. Deep image fingerprint: Towards low budget synthetic image detection and model lineage analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4067–4076, 2024.

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

[27] 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. *arXiv preprint arXiv:2408.09647*, 2024.

[28] 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, pages 5052–5060, 2024.

[29] 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*, pages 28130–28139, 2024.

[30] Milvus Team. Milvus documentation. https://milvus.io/docs/fr, 2025. Accessed: 2025-03-08.

[31] 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*, pages 8695–8704, 2020.

[32] 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*, pages 22445–22455, 2023.

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