Pretrained Image Encoders of Vision-Language and Multimodal Models: An Evaluation on Synthetic Facial Image Detection

Liyue Fan[0000−0002−3819−1098] and Joseph Roberson

University of North Carolina at Charlotte, Charlotte NC 28223, USA
{liyue.fan,jrobe187}@charlotte.edu

Abstract. Accurate detection of synthetically generated data can prevent harmful misuse and facilitate the development of better generative techniques. Prior research has extensively studied the development of machine learning classifiers for synthetic facial image detection. Given the advancement in vision-language and multimodal foundation models and their promising zero-shot performance for downstream tasks, we hypothesize that those models may provide a powerful feature space for accurate and generalizable synthetic data detection. Although recent research studies leveraged multimodal large language models in the context of synthetic image detection, they took a different direction by exploring whether those models can provide natural language explanations for classification outcomes.

In contrast, our study examines the representation space induced by vision-language and multimodal foundation models and investigates its applicability to synthetic facial image detection. The intuition is rooted in the fact that those foundation models are trained on large, a variety of public domain data (i.e., not specifically on facial images or for synthetic image detection tasks) and could capture semantic information in real image data by leveraging knowledge across different modalities. Empirical evaluation with several vision-language and multimodal foundation models is conducted to validate the performance, robustness, and sample efficiency for detecting synthetic facial images generated by a number of diffusion-based methods.

Keywords: Vision-Language Models · Multimodal Models · Image Encoders · Synthetic Image Detection

1 Introduction

Accurate detection of synthetically generated images has important implications. On one hand, it would help regulate the adoption of synthetic data in research and applications, preventing harmful misuse [9, 24]. On the other hand, it would facilitate the development of better data synthesis methods, by creating a symbiosis between generation and detection techniques [10, 13]. In fact, synthetic

data generation has seen rapid improvement in recent years thanks to the advancement of generative machine learning techniques. During the last few years, generation methods for facial image synthesis have evolved from encoder and decoder networks [16], to GANs [6], and recently to using state-of-the-art diffusion models [3]. Rapid evolution and advancement in synthetic facial image generation drives the demand for accurate and generalizable detection methods.

The ability to detect samples from unseen generation methods is a critical challenge in synthetic data detection. Specifically, machine learning based methods trained to detect a specific generation method may fail to detect images generated by other methods [5, 28, 18]. Researchers have argued that supervised detection methods may latch on to artifacts of synthetic samples during training and hence fail to detect new samples from unseen generation methods due to the absence of those artifacts [18]. The study of [18] shows improved generalizability by leveraging the zero-shot ability of a pretrained CLIP ViT model, where a classifier trained on GAN generated images may perform well on images generated by diffusion and autoregressive models.

With vision-language and multimodal foundation models delivering promising results in a range of complex tasks [2, 11], we hypothesize that they can provide a powerful feature space for accurate and generalizable synthetic image detection. Intuitively, foundation models trained on large amounts of public domain data may generalize to detecting different generation methods, as they are not trained to recognize certain objects (e.g., faces) or to detect specific artifacts in synthetic samples. Furthermore, vision-language and multimodal models that leverage knowledge across modalities may capture semantic information in real image samples, providing accurate characterization of real and synthetically generated data. In addition to CLIP ViT models, we consider more recent, advanced foundation models in our study, such as Flamingo [1, 15] and ImageBind [8]. Compared to prior work that leveraged multimodal large language models to provide natural language explanations for synthetic image detection outcomes [26, 27], our focus is on the representation space induced by vision-language and multimodal models and their applicability to synthetic image detection.

Specifically, we adopt pretrained image encoders of vision-language and multimodal models and only train a simple linear classification layer for synthetic image detection on top of sample embeddings. We conduct an empirical evaluation with a range of foundation models and on synthetic face datasets generated by a number of diffusion-based methods. Our results show that frozen image encoders of vision-language and multimodal models are applicable for accurate and generalizable synthetic image detection; they provide comparable and, in several settings, better performance than the supervised baseline in generation method attribution (multi-class classification) and cross-domain detection. Furthermore, recent models demonstrate additional benefits: Flamingo provides comparable robustness against image post-processing as the supervised baseline, without fine-tuning or data augmentation; ImageBind delivers superior detection AUC with limited training samples. Our study demonstrates the promise of leverag-

ing pretrained image encoders in multimodal models for effective and efficient synthetic image detection.

The rest of the paper is organized as follows: Section 2 briefly reviews related work in vision-language and multimodal models and synthetic facial image detection; Section 3 describes the proposed approach and rationale; Section 4 presents the empirical study design and discusses corresponding results; Section 5 concludes the paper and points to future research opportunities.

2 Related Work

2.1 Vision-Language and Multimodal Foundation Models

Foundation models are trained on broad data generally via self-supervision and can be adapted to a wide range of downstream tasks [2]. Publicly available, pretrained vision-language and multimodal foundation models allow deeper understanding of how they operate and greater utilization [14]. For example, CLIP [20] employed large amounts of image-text pairs and embed image and language inputs in a joint space using contrastive learning, which has been shown to exhibit impressive zero-shot performance. Flamingo [1] handles arbitrarily interleaved image and text data, and outperforms several models which have been extensively fine-tuned on task-specific data. ImageBind [8] unifies multiple modalities, including image/video, text, audio, heat map, inertial measurement units, and depth, in one joint embedding space, and achieves state-of-the-art in zero-shot recognition tasks. Recent research shows that foundation models deliver competitive performance in complex tasks such as medical imaging with limited labeled data [11]. A few studies have leveraged large multimodal models to provide natural language explanations for synthetic image detection outcomes [26, 27]. In this study, we leverage the powerful representation space induced by vision-language and multimodal foundation models to recognize a number of diffusion-based, synthetic image generation methods.

2.2 Synthetic Facial Image Detection

The detection of synthetically generated images has been an important problem to study in recent years. Several datasets have been publicly released for research purposes. For example, FaceForensics++ dataset [22], Celeb-DF [16] ,and DFDC [6] are widely adopted datasets for face manipulation detection. They contain large collections of video sequences obtained with several manipulation methods, such as FaceSwap, encoder and decoder models, and StyleGAN. Recently, DiffusionFace [3] provided a face forgery dataset based on diffusion models. It contains synthetic images generated by 11 state-of-the-art diffusion based methods, including unconditional generation methods and conditional generation methods. To detect synthetically generated images, researchers have utilized hand-crafted features [19], observable artifacts [28], and frequency space [7]. One challenge in synthetic image detection is generalizability, i.e., classifiers trained

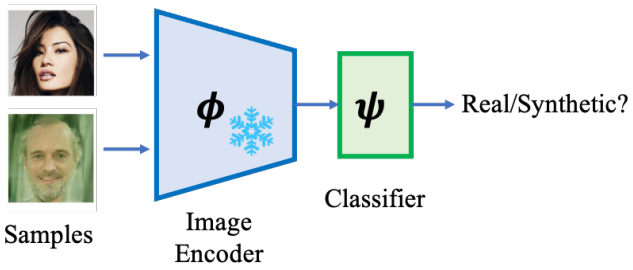


Fig. 1: Synthetic Image Detection with Frozen Image Encoders of Vision-Language and Multimodal Foundation Models

for a specific generative method may fail to detect images generated by other methods [5, 28, 18]. The authors of [18] argued that a pretrained CLIP ViT model has surprisingly good generalization ability, as the feature space is not explicitly trained to distinguish real from synthetic images. Even linear and nearest neighbor classifiers could perform well in detecting synthetic images from unseen generative methods, such as diffusion models. In this study, we investigate this hypothesis in the representation space of recent vision-language and multimodal foundation models, in addition to CLIP ViT.

3 Detecting Synthetic Image using Multimodal Foundation Model

State-of-the-art supervised approaches [28, 23] to detecting synthetically generated data may latch onto artifacts (e.g., frequency patterns) of specific generation methods. As a result, the learned decision boundary may be skewed and may fail to detect synthetic samples lacking those artifacts, e.g., samples generated by new or unseen methods. Prior work [18] has shown that the feature space of the pretrained CLIP ViT image encoder [20], which is not specifically trained to perform real vs. synthetic prediction, results in better generalizability towards detecting unseen generation methods. Specifically, the authors showed simple classifiers trained with the CLIP ViT L/14 encoder on synthetic samples generated by GAN models perform well on samples generated by guided diffusion models. We hypothesize that the feature space of more recent vision-language and multimodal foundation models (such as Flamingo [1] and ImageBind [8]) may further enhance the performance for synthetic image detection.

Our approach is to leverage pretrained image encoders in vision-language and multimodal foundation models and to train only a classifier for image embeddings. As shown in Figure 1, a pretrained image encoder ϕ is employed to extract embeddings from input samples. We propose to add a classification layer ψ on top of the image encoder. Specifically, we implement ψ as a single linear layer with sigmoid activation, as linear probing was shown to be effective and computation friendly [18]. The image encoder ϕ remains frozen during training

and the classifier ψ is trained using binary cross entropy loss:

$$\mathcal{L} = -\sum_{s \in \mathcal{S}} \log(\psi(\phi(s))) - \sum_{r \in \mathcal{R}} \log(1 - \psi(\phi(r))) \quad (1)$$

where $\mathcal{S}$ and $\mathcal{R}$ are synthetic and real samples provided during training.

For *cross-domain* generalization, we assume the classifier ψ has access to real samples $\mathcal{R}$ and synthetic samples $\mathcal{S}$ associated with a single generation method during training. At inference time, the trained classifier may be evaluated on synthetic samples $\mathcal{S}'$ associated with unseen generation methods. This assumption helps examine whether our approach is immune to artifacts of specific generation methods, as opposed to current supervised approaches.

Choice of Image Encoder. Authors of [18] argued that the image encoder ϕ should be trained with different kinds of images so that it can properly embed any new test sample. That motivated their choice of CLIP ViT [20], which was trained on 400M image-text pairs. In this study, we consider multiple variants from the CLIP ViT family to examine the effects of model scale, including the base model with patch size 32 and 16 (i.e., B/32 and B/16) and the deeper large model with patch size 14 (i.e., L/14).

Furthermore, we hypothesize that more recent foundation models (compared to CLIP) which adopt novel paradigms to fuse different data modalities may capture additional semantic information in image modality and can improve synthetic data detection. To that end, we consider Flamingo [1, 15], which employs perceiver samplers and cross-attention blocks to fuse vision and text signals; we also consider ImageBind [8], which aligns the embedding of six different modalities in a common space and requires only image-paired data for training.

4 Experiments

4.1 Experiment Setup

Datasets. We adopt DiffusionFace [3] dataset which provides a comprehensive benchmark for diffusion-based synthetic facial images. DiffusionFace contains 12 subsets, and each subset contains 30,000 image samples. The subsets are

- 1 subset for real images: Multi-Modal-CelebA-HQ [25],
- 5 subsets for samples generated by unconditional generation methods: DDPM, DDIM, PNDM [17], P2 [4], and LDM [21],
- 6 subsets for samples generated by conditional generation methods: text-guided and image-guided generations with Stable Diffusion v1.5 and v2.1, inpainting, and DiffSwap [29].

Models. We leverage a range of vision-language and multimodal models for validation, including OpenAI's CLIP:ViT-B/32, -B/16, and -L/14 [20], open-source reproduction for Flamingo by IDEFICS-9B [15], and ImageBind [8]. We add a linear classification layer to the image encoder in each model and train the

classifier to predict real vs. synthetic samples using Adam optimizer and binary cross entropy loss.

Baselines. As a baseline, we compare the proposed approach with a state-of-the-art supervised method CNNDetection [23], which trains a classification network to predict real vs. synthetic samples. CNNDetection employs a pretrained ResNet-50 model and fine-tunes on real and synthetic samples using Adam optimizer and binary cross entropy. We adopt data augmentation with Gaussian blur and JPEG compression with probability 0.1 during training for improved generalization. In addition, [18] provides another baseline with a linear classifier trained in the representation space of CLIP ViT L/14. Our study incorporates additional models from the same family, i.e., CLIP ViT B/32 and B/16, as well as recent models, i.e., Flamingo and ImageBind.

Implementation. Our approach is implemented in PyTorch. We utilized the official implementation for CNNDetection [23]. Experiments were conducted with batch size 64 and early stopping patience 5, on 2 NVIDIA RTX A4500 GPUs. Note that CNNDetection updates the ResNet backbone during training, which could result in higher computational costs; in contrast, our approach adopts frozen image encoders and trains only the classification layer.

Table 1: Generation Method Attribution Evaluation: best performance in each row in boldface; second best performance in each row underlined.

Metric	CNNDetection	CLIP ViT			Flamingo	ImageBind
		B/32	B/16	L/14		
Top-1 Acc	**0.978** ± 0.000	0.893 ± 0.002	0.917 ± 0.001	0.918 ± 0.000	0.954 ± 0.001	<u>0.957</u> ± 0.000
Top-2 Acc	<u>0.996</u> ± 0.000	0.990 ± 0.001	0.995 ± 0.000	<u>0.996</u> ± 0.000	**1.000** ± 0.000	**1.000** ± 0.000
Top-3 Acc	<u>0.999</u> ± 0.000	0.998 ± 0.000	<u>0.999</u> ± 0.000	<u>0.999</u> ± 0.000	**1.000** ± 0.000	**1.000** ± 0.000
AUC	**1.000** ± 0.000	0.993 ± 0.000	0.995 ± 0.000	0.995 ± 0.000	0.997 ± 0.000	<u>0.998</u> ± 0.000

4.2 Generation Method Attribution

We formulate the method attribution task as identifying the generation method responsible for a given synthetic sample. To that end, we conduct a *multi-class classification* study among all generation methods in the dataset. This study simulates a scenario when annotated synthetic samples from multiple generation methods are available during training; at inference time, we predict if a given sample is real or synthesized by a specific generation method.

Specifically, we adapt the classification layer in the frozen encoder approach as well as in CNNDetection to 12 output classes (including "real") and use cross entropy loss to predict the generation methods of samples. Note that CNNDetection [23] updates both the classification layer and the ResNet-50 backbone with real and synthetic samples during training. We adopt a 80-20 split on each subset for training and testing the classifiers. 5 train runs for each classifier were

conducted with randomly chosen train and test samples. Classification performance is measured by ROC AUC scores (one vs. rest) and top-k accuracy which indicates how often the ground truth label is among the top-k most likely classes predicted.

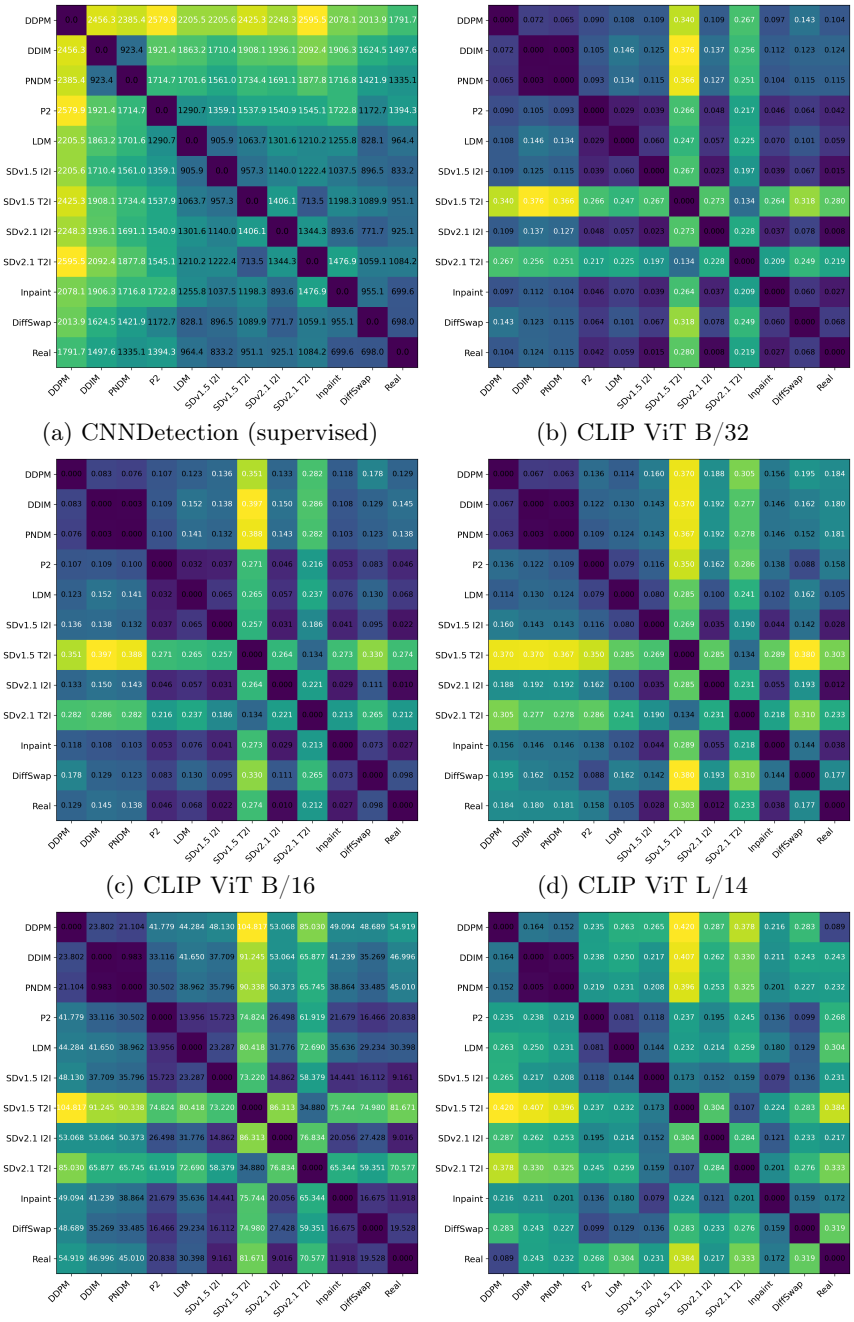
Average performance on test and standard deviation across 5 runs are reported in Table 1. We observe that CNNDetection provides best Top-1 accuracy and AUC score; classifiers built on frozen Flamingo and ImageBind encoders provide comparable performance in those metrics. Furthermore, Flamingo and ImageBind encoders enable perfect Top-2 and -3 accuracies. Among the CLIP family, CLIP ViT B/16 and L/14 perform better than the B/32 model but slightly under-perform Flamingo and ImageBind. Overall, we find that Flamingo and ImageBind provide comparable performance to the supervised CNNDetection, without fine-tuning image encoders on synthetic image samples.

4.3 Fréchet Distance in Representation Space

To better understand the representation space induced by vision-language and multimodal models, we perform a quantitative analysis on image embeddings extracted from the studied image encoders. Specifically, we are interested in the Fréchet distance for samples from different subsets. In a nutshell, the Fréchet distance computes statistics of input data distributions and measures the similarity between two distributions (e.g., DDPM vs. DDIM). A common application of the Fréchet distance is to evaluate synthetic data quality, as in Fréchet Inception distance (FID) [12]; a lower FID score between synthetic and real distributions indicates higher synthetic data quality. However, our experiment studies the representation space induced by image encoders of vision-language and multimodal models, as opposed to that by the Inception model. Furthermore, we apply the Fréchet distance to understand how those frozen image encoders map image samples from various subsets to different distributions (i.e., between real and synthetic samples as well as between different generation methods), as opposed to evaluate synthetic sample quality with respect to real samples only.

For CNNDetection, we extracted image embeddings from the updated ResNet-50 backbone using the corresponding multi-class classification training set, which contains 24,000 images from each subset. For CLIP, Flamingo, and ImageBind, we randomly sampled 24,000 images from each subset and extracted image embeddings from the frozen encoders. Subsequently, we computed the Fréchet distance between embeddings from different subsets for each model, and results are reported in Figure 2.

As in Figure 2a, the Fréchet distance in the supervised CNNDetection representation space is much lower on the diagonal than off-diagonal. It is reasonable as the CNNDetection backbone is trained to maximally "separate" data from different subsets (i.e. classes), leading to distinct distributions in the representation space. In the learned representation space, it can be observed that 4 subsets (DDPM, DDIM, PNDM, and P2) exhibit higher Fréchet distances to other subsets, while the remaining 8 subsets have lower Fréchet distances between each other. Note that real data is most dissimilar to a few unconditional generation



(a) CNNDetection (supervised)

(b) CLIP ViT B/32

(c) CLIP ViT B/16

(d) CLIP ViT L/14

(e) Flamingo

(f) ImageBind

Fig. 2: Fréchet Distance of Image Representations (best in color and when zoomed in)

subsets, such as DDPM, DDIM, and P2, and its Fréchet distances to other synthetic subsets are relatively lower.

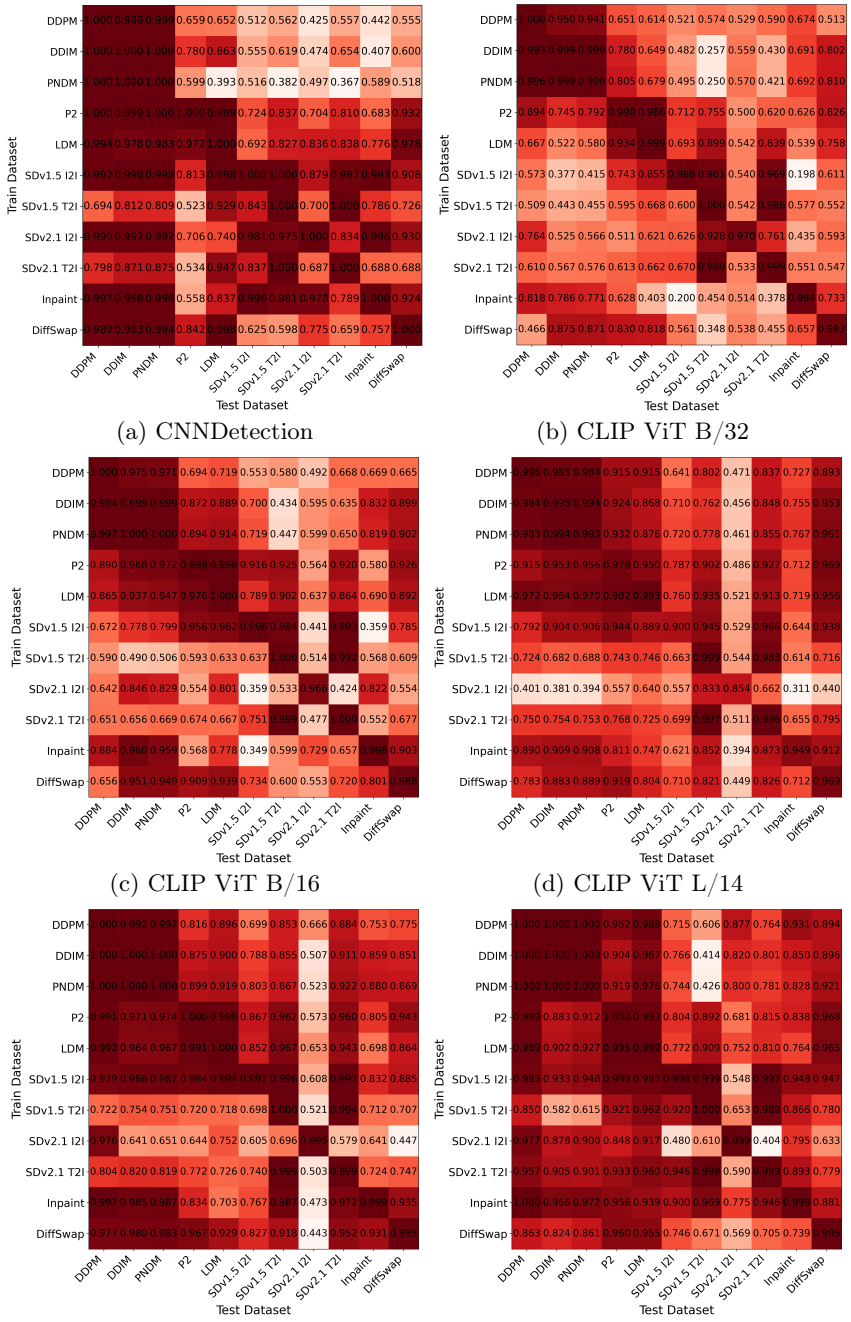
On the other hand, image encoders of the pretrained vision-language and multimodal models capture the distributional difference between different subsets, without being trained on synthetic image detection tasks. Results from CLIP, Flamingo, and ImageBind encoders consistently show that Stable Diffusion text-guided generation (i.e., SDv1.5 T2I and SDv2.1 T2I) exhibit higher Fréchet distances to other image subsets, highlighting the difference in text-guided generation better than the supervised CNNDetection representation. In addition, Flamingo encoder maps DDIM farther away from other subsets (Figure 2e), while CLIP ViT L/14 maps SDv2.1 I2I and DiffSwap farther away from other subsets (Figure 2d). Neither Flamingo nor CLIP encoders fully highlight the difference between real data and synthetic subsets. The ImageBind representation best captures the distributional differences between real and synthetically generated data, leading to higher Fréchet distances between real and almost all synthetic subsets (Figure 2f). Furthermore, difference between different synthetic subsets is better highlighted in the ImageBind representation space, compared to CLIP, Flamingo, and CNNDetection. We hypothesize that ImageBind representation space captures subtle differences between subsets as the image encoder learns to preserve semantic information across multiple data modalities. These results demonstrate the generalizability of the multi-modal ImageBind model to the task of generation method attribution.

4.4 Cross-domain Detection

We conduct a cross-domain classification study to understand if classifiers trained to recognize *specific* generation methods could be applied to *unseen* generation methods. This setup simulates a realistic scenario for detecting synthetic samples from new, unknown generation methods. Specifically, we trained a *binary* classifier for real vs. synthetic detection for each generation method and tested the classifier on every other generation method. Similar to Section 4.2, we adopt a 80-20 split on each subset for training and testing the classifiers. Test AUC results are reported in Figure 3.

Clearly, classifiers trained to detect specific synthetic image types perform well on within-domain tests, i.e., high AUCs on the diagonal. The baseline CNNDetection classifiers trained on DDPM, DDIM, and PNDM do not generalize well to unseen generation methods, leading to low test AUC on other subsets. We observe various degrees of cross-domain generalization for vision-language and multimodal models, and ImageBind and Flamingo representations achieve better cross-domain generalization compared to CLIP ViT models. Among CLIP ViT variants, the deeper L/14 model outperforms B/32 and B/16 models.

In addition, we observe that in the feature space of vision-language and multimodal models, classifiers trained on other generation methods perform worse on detecting SD v2.1 I2I samples. The effect is less severe in ImageBind results 3f. On the other hand, CNNDetection results also showed that SD v2.1 I2I and Inpaint are more difficult to detect as *unseen*. Overall, ImageBind and



(a) CNNDetection (b) CLIP ViT B/32

(c) CLIP ViT B/16 (d) CLIP ViT L/14

(e) Flamingo (f) ImageBind

Fig. 3: Cross-Domain Detection Results in AUC (best in color and when zoomed in)

Flamingo representations enable better cross-domain generalization compared to CLIP ViT models and CNNDetection. We hypothesize that adopting the frozen image encoders in ImageBind and Flamingo may help mitigate overfitting effects, whereas the CNNDetection representation space may overfit to samples from the specific generation method during training.

4.5 Effects of Model Scale

We aim to understand the effects of model scale in the same model family, by comparing three models in OpenAI's CLIP ViT. The number of parameters in each model ranges from ∼149 million (CLIP ViT B/32), ∼151 million (CLIP ViT B/16), to ∼427 million (CLIP ViT L/14). The performance comparison is presented in Table 1, Figure 2, and Figure 3.

From Table 1, we observe that the representation space of all CLIP models can effectively capture the difference between real and synthetic samples from various generation methods, with B/16 and L/14 models providing better results than B/32. In Figure 2, CLIP ViT L/14 model highlights the difference between real and synthetic samples more, compared to B/32 and B/16 models. CLIP ViT L/14 provides much better performance in cross-domain generalization as shown in Figure 3. We hypothesize that the large size of CLIP ViT L/14 may allow for better modeling of the real data distribution, thus enabling better distinction of different datasets and cross-domain generalization. More principled, in-depth study on the representation capacity of different model scales may be conducted by future research.

Table 2: Generation Method Attribution Robustness : best performance in each row in boldface; second best performance in each row underlined.

Setting/Metric		CNNDetection	CLIP ViT L/14	Flamingo	ImageBind
Resize (/16)	Top-1 Acc	**0.267** ± 0.012	<u>0.200</u> ± 0.002	<u>0.200</u> ± 0.001	0.098 ± 0.000
	Top-2 Acc	**0.455** ± 0.023	0.356 ± 0.002	<u>0.378</u> ± 0.001	0.209 ± 0.002
	Top-3 Acc	**0.596** ± 0.026	0.473 ± 0.002	<u>0.495</u> ± 0.001	0.327 ± 0.004
	AUC	**0.764** ± 0.016	0.695 ± 0.002	<u>0.711</u> ± 0.001	0.648 ± 0.001
JPEG ($q = 10$)	Top-1 Acc	**0.267** ± 0.012	0.207 ± 0.002	<u>0.235</u> ± 0.001	0.108 ± 0.000
	Top-2 Acc	**0.455** ± 0.023	0.343 ± 0.003	<u>0.378</u> ± 0.001	0.233 ± 0.003
	Top-3 Acc	**0.595** ± 0.031	0.464 ± 0.002	<u>0.495</u> ± 0.001	0.343 ± 0.002
	AUC	**0.763** ± 0.016	0.700 ± 0.003	<u>0.711</u> ± 0.001	0.647 ± 0.001
Gaussian Blur ($r = 22$)	Top-1 Acc	<u>0.225</u> ± 0.030	0.209 ± 0.000	**0.272** ± 0.002	0.098 ± 0.001
	Top-2 Acc	**0.402** ± 0.041	<u>0.338</u> ± 0.000	**0.402** ± 0.002	0.240 ± 0.003
	Top-3 Acc	**0.545** ± 0.042	0.435 ± 0.001	<u>0.514</u> ± 0.001	0.365 ± 0.000
	AUC	**0.729** ± 0.025	0.700 ± 0.002	<u>0.725</u> ± 0.002	0.653 ± 0.001

4.6 Robustness against Post-processing

As real-world synthetic images may not appear in pristine form, it is important to understand if classifiers may fail to detect synthetic samples if image post-processing have been applied to evade detection. To that end, we transform test samples in generation method attribution and evaluate model performance on transformed samples. We consider common image post-processing that could inflict non-trivial quality loss, i.e., resizing (with target height and width 16 times smaller than original values), JPEG compression (with quality set to 10), and Gaussian blur (with radius set to 22). Similar to Section 4.2, we adopt 80-20 train-test split and conduct 5 runs for each approach. Average and standard deviation results on transformed test samples are reported in Table 2.

It can be seen that all transformations make classification more difficult for both the baseline and vision-language and multimodal models, compared to clean test samples. For instance, highest Top-1 Accuracy is at 0.267 and highest AUC is at 0.764 across all classifiers and transformations. Although the baseline CNNDetection has been trained with data augmentation using Gaussian blur and JPEG compression, its performance degrades when test samples present higher amounts of distortion than training augmentation (such as larger radius for Gaussian blur). Although the results show that classifiers based on vision-language and multimodal models are affected by image post-processing, they provide comparable performance to the CNNDetection baseline without training the image encoder or with data augmentation. Flamingo outperforms CLIP ViT L/14 and ImageBind across different transformations, providing a more robust representation space against the studied image post-processing.

4.7 Effects of Limited Training Samples

Last but not least, large amounts of synthetic samples for training may be challenging to obtain in practice, especially for method attribution tasks where training requires samples from multiple generation methods. As a result, it is important to develop classifiers that are sample-efficient and can generalize well from limited training data. To that end, we limit the number of training samples in multi-class classification (e.g. to as low as 5 samples per class) and evaluate the trained classifier on the original test split, i.e., 20% samples from each subset/class. Average and standard deviation results over 5 different train runs are reported in Table 3.

We observe that vision-language and multimodal models demonstrate better performance than the baseline when training samples are very limited (≤ 10 per class), enabling more accurate classification with high AUC. Most notably, ImageBind achieves average test AUC 0.944 with only 5 training samples per class. When 30 samples per class are available during training, all models including the CNNDetection baseline achieve ≥ 0.9 average test AUC. ImageBind consistently provides the highest performance when varying the number of training samples. Based on our results, the multimodal model appears most suitable in synthetic image detection settings with limited training data.

Table 3: Generation Method Attribution (Test AUC) with Limited Training Samples: best performance in each row in boldface; second best performance in each row underlined.

#samples per class	CNNDetection	CLIP ViT L/14	Flamingo	ImageBind
5	0.786 ± 0.015	$\underline{0.808} \pm 0.008$	0.803 ± 0.027	$\mathbf{0.944} \pm 0.012$
10	0.855 ± 0.005	$\underline{0.861} \pm 0.009$	0.854 ± 0.010	$\mathbf{0.961} \pm 0.004$
15	$\underline{0.883} \pm 0.011$	0.880 ± 0.004	0.879 ± 0.015	$\mathbf{0.970} \pm 0.003$
30	$\underline{0.924} \pm 0.005$	0.908 ± 0.001	0.911 ± 0.003	$\mathbf{0.978} \pm 0.004$

5 Conclusion and Discussion

We examined a number of vision-language and multimodal foundation models in the application of detecting synthetic facial images generated by a range of diffusion based methods. We extended the zero-shot approach based on CLIP ViT models in [18] to more recent vision-language and multimodal models such as Flamingo and ImageBind. We found that compared to the supervised CNNDetection approach, Flamingo and ImageBind image encoders provide comparable or, in several settings, more accurate detection of multiple generation methods, unseen generation methods, and against a number of image transformations at inference time. Furthermore, the multimodal ImageBind model is highly efficient in the number of training samples, providing accurate classification with as low as 5 samples per generation method. We conclude that the representation space of vision-language and multimodal foundation models may be leveraged for practical and effective detection of synthetic facial images, without costly fine-tuning large backbones or training networks from scratch.

One immediate improvement to our study is that future work could conduct more training runs (5 runs were conducted in this study) with limited training samples. With 5 to 30 samples per class, classifier performance may be influenced by the selected training samples; averaging over a large number of training runs could provide more accurate estimates. Furthermore, future work may conduct a comprehensive evaluation with multiple real data sources (such as FFHQ) and more generation methods (such as DALL-E and Midjourney). Incorporating multiple datasets will help understand whether classifiers simply memorize specific signatures or artifacts of the real source rather than learning authenticity. Moreover, it is important to further study the representation space of vision-language and multimodal foundation models, in order to fully analyze their potential benefits in performance, robustness, and sample efficiency. Our study presented a quantifiable evaluation on Fréchet distance between sample embeddings, and future research may consider additional quantifiable approaches and metrics.

6 Acknowledgments

This work was supported by the National Science Foundation under grant CNS-2027114. The opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Bommasani, R.: On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258 (2021)
3. Chen, Z., Sun, K., Zhou, Z., Lin, X., Sun, X., Cao, L., Ji, R.: Diffusionface: Towards a comprehensive dataset for diffusion-based face forgery analysis. arXiv preprint arXiv:2403.18471 (2024)
4. Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., Yoon, S.: Perception prioritized training of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11472–11481 (2022)
5. Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., Verdoliva, L.: Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510 (2018)
6. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020)
7. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: International conference on machine learning. pp. 3247–3258. PMLR (2020)
8. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15180–15190 (2023)
9. Giuffrè, M., Shung, D.L.: Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ digital medicine **6**(1), 186 (2023)
10. Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G., Verdoliva, L.: Are gan generated images easy to detect? a critical analysis of the state-of-the-art. arXiv preprint arXiv:2104.02617 (2021)
11. Han, T., Adams, L.C., Nebelung, S., Kather, J.N., Bressem, K.K., Truhn, D.: Multimodal large language models are generalist medical image interpreters. medRxiv pp. 2023–12 (2023)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
13. Hou, Y., Guo, Q., Huang, Y., Xie, X., Ma, L., Zhao, J.: Evading deepfake detectors via adversarial statistical consistency. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12271–12280 (2023)

14. Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A.K., Bankston, K., Biderman, S., Bogen, M., et al.: Position: On the societal impact of open foundation models. In: Forty-First International Conference on Machine Learning (2024)
15. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A., Kiela, D., et al.: Obelics: An open web-scale filtered dataset of interleaved image-text documents. Advances in Neural Information Processing Systems **36**, 71683–71702 (2023)
16. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3207–3216 (2020)
17. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. arXiv preprint arXiv:2202.09778 (2022)
18. Ojha, U., Li, Y., Lee, Y.J.: 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)
19. Popescu, A.C., Farid, H.: Exposing digital forgeries by detecting traces of resampling. IEEE Transactions on signal processing **53**(2), 758–767 (2005)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
21. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
22. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1–11 (2019)
23. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8695–8704 (2020)
24. Wu, Y., Yang, Z., Shen, Y., Backes, M., Zhang, Y.: Synthetic artifact auditing: Tracing llm-generated synthetic data usage in downstream applications. arXiv preprint arXiv:2502.00808 (2025)
25. Xia, W., Yang, Y., Xue, J.H., Wu, B.: Tedigan: Text-guided diverse face image generation and manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2256–2265 (2021)
26. Ye, J., Zhou, B., Huang, Z., Zhang, J., Bai, T., Kang, H., He, J., Lin, H., Wang, Z., Wu, T., et al.: Loki: A comprehensive synthetic data detection benchmark using large multimodal models. arXiv preprint arXiv:2410.09732 (2024)
27. Yu, P., Fei, J., Gao, H., Feng, X., Xia, Z., Chang, C.H.: Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection. arXiv preprint arXiv:2503.14853 (2025)
28. Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. In: 2019 IEEE international workshop on information forensics and security (WIFS). pp. 1–6. IEEE (2019)
29. Zhao, W., Rao, Y., Shi, W., Liu, Z., Zhou, J., Lu, J.: Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8568–8577 (2023)