ANTIFAKEPROMPT: PROMPT-TUNED VISION-LANGUAGE MODELS ARE FAKE IMAGE DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep generative models can create remarkably photorealistic fake images while raising concerns about misinformation and copyright infringement, known as deepfake threats. Deepfake detection technique is developed to distinguish between real and fake images, where the existing methods typically learn classifiers in the image domain or various feature domains. However, the generalizability of deepfake detection against emerging and more advanced generative models remains challenging. In this paper, being inspired by the zero-shot advantages of Vision-Language Models (VLMs), we propose a novel approach using VLMs (e.g. InstructBLIP) and prompt tuning techniques to improve the deepfake detection accuracy over unseen data. We formulate deepfake detection as a visual question answering problem, and tune soft prompts for InstructBLIP to answer the real/fake information of a query image. We conduct full-spectrum experiments on datasets from 3 held-in and 16 held-out generative models, covering modern text-to-image generation, image editing and adversarial image attacks. Results demonstrate that (1) the deepfake detection accuracy can be significantly and consistently improved (from 61.79% to 92.72%, in average accuracy over unseen data) using pretrained vision-language models with prompt tuning; (2) our superior performance is at less cost of training data and trainable parameters, resulting in an effective and efficient solution for deepfake detection. Data, code, models will be open-sourced.

1 INTRODUCTION

In recent years, we have witnessed the magic leap upon the development of generative models, where the cutting-edge models such as Stable Diffusion (Rombach et al., 2022), DALLE-2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022) and DALLE-3 (OpenAI, 2023) have become capable of producing high-quality images, ranging from beautiful artworks to incredibly realistic images.

However, the progress of such image synthesis technique, which is called “deepfake", poses real threats to our society, as some realistic fake images could be produced to deceive people and spread false information. For example, images of the war in Ukraine could be generated with having the false or misleading information and may be used for propaganda [1]. More concerningly, some of these created images may be falsely claimed as works of photographers or artists, potentially leading to copyright infringements and their misuse in commercial contexts. As BBC reported, some fake artworks generated by text-to-image generation models won first place in an art competition, which harming the fairness of the contest[2]. To protect against these threats arising from deepfake content, the use of effective deepfake detection techniques becomes crucial. These techniques help distinguish real content from manipulated images, serving as a vital defense against deception and safeguarding intellectual property rights in the digital era.

One straightforward deepfake detection prototype is to train a classifier to distinguish between real and fake images (Wang et al., 2020; Guarnera et al., 2023; Yu et al., 2019). However, along with the rapid development of generative models, this approach often struggles with overfitting thus leading to poor performance on unseen data from emerging generators. To overcome this limitation,

[1] https://techcrunch.com/2022/08/12/a-startup-wants-to-democratize-the-tech-behind-dall-e-2-consequences-be-damned/

[2] https://www.bbc.com/news/technology-62788725

researchers are exploring more general features in deepfake images, such as frequency maps of the images (Zhang et al., 2019b). Additionally, some innovative methods are not only based on visual features. For example, DE-FAKE (Sha et al., 2022) harnesses the power of large language models and trains a classifier that conditions on both visual and textual information.

Despite years of development in deepfake detection techniques, challenges persist. First, most previous works such as (Wang et al., 2020) have concentrated on Generative Adversarial Networks (GANs), which may not effectively address the latest diffusion-based generative models including Stable Diffusion (Rombach et al., 2022), Stable Diffusion-XL (Podell et al., 2023), DALLE-2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), DeepFloyd IF (StabilityAI, 2023). Second, generalizability remains a significant challenge. Classifiers trained on images generated by one model tend to perform poorly when being tested on images from different generative models, especially from more emerging and advanced ones.

To address the aforementioned challenges and to utilize the strong generality of LLMs, we have harnessed the zero-shot capabilities of pretrained vision-language models (Li et al., 2022; 2023; Zhu et al., 2023; Liu et al., 2023; Dai et al., 2023) to capture more general instruction-aware features from images, enhancing the transferability of our deepfake detector. We formulate the deepfake detection problem as a Visual Question Answering (VQA) task, asking the model with the question "Is this photo real?", to tackle this challenge. However, directly asking questions for a pretrained VLM may not lead to effective answers, considering either the query images or questions are unseen during VLM training. We therefore use prompt tuning to boost the performance. Without loss of generality, we build implementation on the recent state-of-the-art VLM, InstructBLIP (Dai et al., 2023). Specifically, we insert a "pseudo-word" into the prompt and optimize the corresponding word embedding in the model for correctly answering “Yes" and “No" to the question for real and fake images on training data, respectively. This approach not only significantly reduces training costs but also substantially improves performance on both held-in and held-out testing datasets from a full spectrum of generative models. From the perspective of instruction tuning, we realized that there are many good answers, all waiting for a good question. In summary, our paper makes the following key contributions:

1. We pioneer to leverage pretrained vision-language models to solve the deepfake detection problem. We are the first to formulate the problem as a VQA scenario, asking the model to distinguish between real and fake images. Additionally, we employ soft prompt tuning techniques to optimize for the most effective question to the VLMs, and leverage their zero-shot generalizability on unseen data produced by held-out generative models.
2. Our detector consistently outperforms the recent baseline methods proposed in (Wang et al., 2020; Sha et al., 2022; Wang et al., 2023; Wu et al., 2023; Ricker et al., 2022; Le & Woo, 2023) over held-out datasets generated by a full spectrum of generator categories. Our superior performance and generalizability benefit from the nature of pretrained VLMs, and at less cost of training data and trainable parameters.

2 Related Work

2.1 Visual Generative Models

The recent advance of deep generative models can be broadly categorized into two main types: Generative-Adversarial-Networks-based (GAN-based) models and diffusion-based models. Within the realm of GAN-based model, notable progress has been made. Starting from GAN (Goodfellow et al., 2014), SA-GAN (Zhang et al., 2019a) and BigGAN (Brock et al., 2018) contributed to the enhancement of training stability and the generation of diverse images with higher resolution. Subsequently, StyleGAN (Karras et al., 2019) and its successors (Karras et al., 2020; 2021) have allowed for finer control over the stylistic attributes of the generated images while maintaining high image quality. Building upon StyleGAN-3 (Karras et al., 2021) and ProjectGAN (Sauer et al., 2021), StyleGAN-XL (Sauer et al., 2022b) is able to generate 1024×1024 images with even lower Fréchet Inception Distance (FID) scores and higher Inception Scores (IS) , w.r.t. all its predecessors.

In regard to the diffusion-based models, starting from DDPM (Ho et al., 2020), DDIM (Song et al., 2020) speeds up the generating process by relaxing the constraint of Markov Chain towards forward and backward processes. Latent Diffusion (Rombach et al., 2021) and Stable Diffusion (Rombach et al., 2022) further shift the diffusion process to latent space, granting user controls over the models; thus, it flexibly enables the text-to-image generation through diffusion-based models. Building upon this foundation, several seccessors (e.g. SDXL (Podell et al., 2023), DeepFloyd IF (StabilityAI, 2023), Imagen (Saharia et al., 2022), Dalle-2 (Ramesh et al., 2022), and Dalle-3 (OpenAI, 2023)) further refine the text comprehension capabilities of diffusion-based models, enabling them to create images that better align with input texts.

Apart from text-to-image generation, image editing tasks, such as inpainting and super resolution, are also widely-used applications of generative models. Notably, (Suvorov et al., 2022; Rombach et al., 2022; Liu et al., 2020) have demonstrated exceptional performances in the domain of image inpainting, while (Lee & Jin, 2022; Rombach et al., 2022; Chen et al., 2021) are known for their remarkable performances in image super resolution.

Without the loss of representativeness, we select a diverse set of generative models (namely SD2, SDXL, IF, Dalle-2, SGXL, ControlNet, LaMa, LTE, SD2 inpainting model, and SD2 super resolution model) to cover the full spectrum of generation tasks, and generate corresponding fake images for conducting our experiments.

2.2 Deepfake Detection Methods

Recent advances in detection methods have focused on training detectors capable of identifying artifacts specific to certain types of generative models. For example, (Wang et al., 2020; Nataraj et al., 2019; Yu et al., 2019; Ricker et al., 2022; Wang et al., 2023; Wu et al., 2023; Ma et al., 2023; Lorenz et al., 2023) leverage artifacts from synthesized images generated by GANs or diffusion models, (Zhang et al., 2019b; Giudice et al., 2021; He et al., 2021) concentrate on artifacts in the frequency domain, and (Le & Woo, 2023) also study the detection of low-quality or low-resolution fake images. These methods have reported outstanding performance on images generated by the seen models, but they often suffer from significant drops in performance when being applied to unseen datasets. Therefore, we aim to propose a general detector that can demonstrate exceptional performance on both in-domain and out-of-domain datasets.

2.3 Vision-Language Models and Visual Question Answering

With the impressive success of Large language models (LLM) (Chung et al., 2022; Touvron et al., 2023), recent studies work on Vision-Language Models (VLMs) (Li et al., 2022; 2023; Zhu et al., 2023; Liu et al., 2023; Ye et al., 2023; Dai et al., 2023) to improve multimodal comprehension and generation through utilizing the strong generalizability of LLMs. These models takes advantage of cross-modal transfer, allowing knowledge to be shared between language and multimodal domains. BLIP-2 (Li et al., 2023) employing a Flan-T5 (Chung et al., 2022) with a Q-former to efficiently align the visual features with language model. MiniGPT-4 (Zhu et al., 2023) employs the pretrained visual encoder and Q-Former as used in BLIP-2, but chooses Vicuna (Chiang et al., 2023) as the LLM and performs training using ChatGPT [3]-generated image captions instead of the BLIP-2 training data. InstructBLIP (Dai et al., 2023) also utilizes the pretrained visual encoder and Q-former from BLIP-2, with Vicuna/Flan-T5 as pretrained LLM, but performs instruction tuning on Q-former using a variety of vision-language tasks and datasets. LLaVA (Liu et al., 2023) projects the output of a visual encoder as input to a LLaMA/Vinuca LLM with a linear layer, and finetunes the LLM on vision-language conversational data generated by GPT-4 (OpenAI, 2023) and ChatGPT. mPLUG-owl (Ye et al., 2023) finetunes a low-rank adaption (Hu et al., 2021) module on a LLaMA (Touvron et al., 2023) model using both text instruction data and vision-language instruction data from LLaVA.

Among the vision-language tasks, visual question answering (VQA) problem is one of the most general and practical tasks because of its flexibility in terms of the questions. For training the models for VQA problems, lots of datasets (Goyal et al., 2017; Gurari et al., 2018; Marino et al., 2019; Schwenk et al., 2022; Mishra et al., 2019; Singh et al., 2019) have been proposed. VQAv2 (Goyal et al., 2017) and VizWiz (Gurari et al., 2018) collect images, questions, as well as the corresponding answers for studying visual understanding. OKVQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022) propose visual question-answer pairs with external knowledge (e.g. Wikipedia). OCR-VQA

[3] https://chat.openai.com/

(Mishra et al., 2019) and TextVQA (Singh et al., 2019) introduce images and questions that require reasoning about text to answer. To solve the VQA problem, TextVQA propose a method called LoRRA, which reads the text in the image and predicts the answer which might be a deduction based on the text and the image or composed of the strings found in the image. On top of that, the aforementioned VLMs can also be potential solutions, as they have strong multimodal comprehension and generality.

Given the remarkable multimodal capabilities of VLMs, we have harnessed their potential to address the deepfake detection challenge. Therefore, we formulate the deepfake detection method as a VQA problem to take advantage of the capabilities of these VLMs.

Furthermore, as (Chen et al., 2023; Zou et al., 2023; Deng et al., 2022; Zhang et al., 2022) concluded, prompt tuning offers an approach to enabling Langnuage Models (LMs) to better understand user-provided concepts, and improves the alignments between generated images and the input prompts when applied to text-to-image generative models (Wen et al., 2023; Gal et al., 2022). Inspired by these findings, we apply prompt tuning atop InstructBLIP to optimize an instruction that can more accurately describe the idea of differentiating real and fake images, resulting in better performance.

3 AntifakePrompt

3.1 Problem Formulation

In order to take advantage of the vision-language model, we formulate the deepfake detection problem as a visual question answering (VQA) problem. In this framework, the input consists of a query image $\mathbf{I}$ that needs to be classified as real or fake and a question prompt q. The prompt can be either a preset question (e.g., "Is this photo real?") or a tunable question that includes the pseudo-word S_*. The output of this framework corresponds to the answer texts y. While y in principle can be any texts, we constrain it to two options: “Yes” and “No” during testing, aligning with the answer ground truth to the original binary classification problem. We choose the option with a higher probability from the VLM as the answer, where the model capability is evaluated by classification accuracy.

In summary, the deepfake detection task can be formulate as a VQA task, which is defined as:

$$\mathcal{M}(\mathbf{I}, q) \mapsto y \tag{1}$$

where $\mathcal{M}$ is an VLM and we adopt InstructBLIP (the recent state-of-the-art) for building our method, and the text output $y \in \{\text{“Yes”}, \text{“No”}\}$ corresponds to the binary results of deepfake detection.

3.2 Prompt tuning on InstructBLIP

As discussed in (Dai et al., 2023), the prompt plays an essential role in VQA problem, and asking the preset question leads to ineffective performance on unseen data. Therefore, we employ soft prompt tuning on InstructBLIP (Dai et al., 2023) following the procedure below.

Within InstructBLIP, two components receive the prompt as input: Q-Former and the Large Language Model (LLM). As shown in Figure 1, the prompt first gets tokenized and embedded, and then is fed into Q-Former and the LLM in parallel. We introduce a pseudo-word S_* into the prompt, which serves as the target for soft prompt tuning. Specifically, we adopt the question template, “Is this photo real?” and append the pseudo-word to the end of the prompt, resulting in the modified prompt q_*: “Is this photo real S_*?”. As the prompt has been decided, we give the output label $\hat{y} =$ “Yes” for real images and $\hat{y} =$“No” for fake images in order to perform soft prompt tuning.

We freeze all parts of the model except for the word embedding v of the pseudo-word S_*, which is randomly initialized. Then we optimize the word embedding v_* of the pseudo-word over a training set of triplet $\{\mathbf{I}, q_*, \hat{y}\}$ with respect to the language modeling loss, expecting the VLM output y to be the label $\hat{y}$. Hence, our optimization goal can be defined as:

$$\widetilde{S_*} = \arg\min_{S_*} \mathbb{E}_{(\mathbf{I},\hat{y})} \mathcal{L}(\mathcal{M}(\mathbf{I}, \text{“Is this image real } S_*\text{”}), \hat{y}) \tag{2}$$

where $\mathcal{L}$ is the language modeling loss function. Since we actually optimize the embedding v_* for the pseudo-word S_*, with noting the concatenation of embeddings for the original prompt (i.e. “'Is

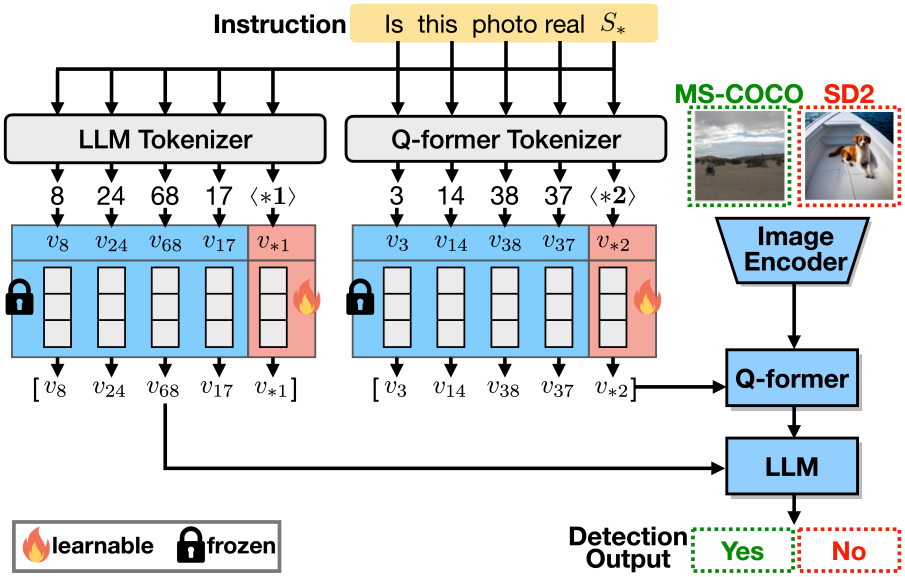
Instruction
Is this photo real S_*
LLM Tokenizer
Q-former Tokenizer
8 24 68 17 ⟨*1⟩
3 14 38 37 ⟨*2⟩
v_8 v_{24} v_{68} v_{17} v_{*1}
v_3 v_{14} v_{38} v_{37} v_{*2}
[v_8 v_{24} v_{68} v_{17} v_{*1}]
[v_3 v_{14} v_{38} v_{37} v_{*2}]
MS-COCO
SD2
Image Encoder
Q-former
LLM
learnable
frozen
Detection Output
Yes
No

Figure 1: **Prompt tuning on InstructBLIP (Dai et al., 2023) for deepfake detector training**. An instruction containing a pseudo-word S_* is first converted into tokens. These tokens are converted to continuous vector representations (the “embeddings”, v). Then, the embedding vectors are fed into Q-former and LLM with the image features extracted by the image encoder. Finally, the embedding vectors v_{*1} and v_{*2} are optimized using language modeling loss, expecting the output to be “Yes” for real images and “No” for fake images.

this photo real”) to be v_p, the equation can be rewritten as:

$$\widetilde{v_*} = \arg\min_{v_*} \mathbb{E}_{(\mathbf{I},\hat{y})} \mathcal{L}(\mathcal{M}(\mathbf{I}, [v_p, v_*]), \hat{y}) \tag{3}$$

As Figure 1 shows, it is crucial to highlight that the pseudo-word embedding fed into Q-Former v_{*1} differs from that fed into the LLM v_{*2}, and we optimize these two embeddings independently. The dimensions of v_{*1} and v_{*2} are 768 and 4096 respectively, so the number of trainable parameters is 4864 in total. Compared to 23 million trainable parameters from ResNet-50 (He et al., 2016) of (Wang et al., 2020) and 11 million trainable parameters from ResNet-18 (He et al., 2016) of DE-FAKE (Sha et al., 2022), our method demonstrates superior cost-efficiency.

Implementation details. We use the LAVIS library[4] for implementation, training, and evaluation. To avoid the out-of-memory issue on small GPUs, we choose Vicuna-7B (Chiang et al., 2023), a decoder-only Transformer instruction-tuned from LLaMA (Touvron et al., 2023), as our LLM. During prompt tuning, we initialize the model from instruction-tuned checkpoint provided by LAVIS, and only finetune the word embedding of the pseudo-word while keeping all the other parts of the model frozen. All models are prompt-tuned with a maximum of 10 epochs. We use AdamW (Loshchilov & Hutter, 2017) optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, batch size 6 and a weight decay 0.05. We initially set the learning rate to 10^{-8}, and apply a cosine decay with a minimum learning rate of 0. All models are trained utilizing 4 NVIDIA RTX 3090 GPUs and completed within 10 hours. In terms of image preprocessing, all images are initially resized to have a length of 224 pixels on the shorter side while maintaining their original aspect ratio. During the training phase, random cropping is applied to achieve a final size of 224 × 224 pixels, while images are center cropped to a final size of 224 × 224 pixels in the testing phase.

4 Experiments

4.1 Setup

Datasets We use Microsoft COCO (MS COCO) (Lin et al., 2014) dataset and Flickr30k (Young et al., 2014) dataset. Being widely utilized as benchmarks of object detection and captioning task,

[4] `https://github.com/salesforce/LAVIS`

Table 1: **Held-in and held-out deepfake detection accuracies**. Experiments are conducted on 2 real and 19 fake datasets, including 3 attacked ones. The accuracies of real and fake out-of-domain datasets are highlighted in green and red, respectively, and average accuracies are highlighted in blue. The accuracies of real and fake in-domain datasets are in grey with lighter green and lighter red background color. The best performances are denoted in **bold**. We mark the training set of InstructBLIP to be "-" to indicate that we use the pretrained model (Dai et al., 2023) without additional training set.

Methods	**Training set**	**No. of param.**	**MS COCO**	**Flickr**	**SD2**	**SDXL**	**IF**	**DALLE-2**	**SGXL**
Wang 2020	ImageNet vs. ProGAN	23.51M	96.87	96.67	0.17	0.17	19.17	3.40	79.30
DE-FAKE	MS COCO vs. SD2	308.02M	85.97	90.67	97.10	90.50	**99.20**	68.97	56.90
DIRE	LSUN B. vs. ADM	23.51M	81.77	77.53	3.83	18.17	6.93	2.13	45.27
LASTED	LSUN, Danbooru vs. ProGAN, SD1.5	625.63M	75.47	76.33	58.69	51.33	57.99	57.96	64.39
J. Ricker 2022	LSUN B. vs. 5 GANs, 5 DMs	23.51M	95.60	95.80	81.10	**99.70**	92.65	52.10	**100.00**
QAD	7 face swapping datasets	2.56K	65.93	65.47	37.93	45.60	35.30	39.47	30.23
InstructBLIP	-	188.84M	**98.93**	**99.63**	40.27	23.07	20.63	41.77	69.53
InstructBLIP + LoRA	MS COCO vs. SD2	4.19M	95.73	91.83	**98.03**	96.33	86.60	**99.57**	97.67
AntifakePrompt	MS COCO vs. SD2	**4.86K**	95.37	91.00	97.83	97.27	89.73	**99.57**	99.97
	MS COCO vs. SD2+LaMa	**4.86K**	90.83	81.04	97.10	97.10	88.37	99.07	99.93

Methods	**Training set**	**No. of param.**	**GLIDE**	**ControlNet**	**Deeper-Forensics**	**DFDC**	**FaceForensics++**	**Inpainting**	
								LaMa	**SD2**
Wang 2020	ImageNet vs. ProGAN	23.51M	17.23	11.43	0.30	0.00	5.23	7.53	0.17
DE-FAKE	MS COCO vs. SD2	308.02M	76.50	63.97	86.97	56.13	78.90	13.03	16.00
DIRE	LSUN B. vs. ADM	23.51M	4.63	9.90	0.27	60.13	25.50	13.23	11.37
LASTED	LSUN, Danbooru vs. ProGAN, SD1.5	625.63M	54.46	50.70	86.38	70.19	70.69	60.53	56.96
J. Ricker 2022	LSUN B. vs. 5 GANs, 5 DMs	23.51M	83.80	75.50	14.20	46.90	20.30	**64.30**	59.10
QAD	7 face swapping datasets	2.56K	55.80	36.37	63.20	77.20	93.93	36.83	34.27
InstructBLIP	-	188.84M	37.97	33.97	13.83	14.07	44.20	10.90	44.23
InstructBLIP + LoRA	MS COCO vs. SD2	4.19M	95.90	92.87	**98.80**	90.03	94.70	59.50	**93.03**
AntifakePrompt	MS COCO vs. SD2	**4.86K**	99.17	91.47	97.90	**100.00**	97.43	39.03	85.20
	MS COCO vs. SD2+LaMa	**4.86K**	**99.73**	**93.27**	97.77	**100.00**	**98.30**	58.53	90.70

Methods	**Training set**	**No. of param.**	**Super Res.**			**Attack**		**Average**
			LTE	**SD2**	**Adver.**	**Backdoor**	**Data Poisoning**	
Wang 2020	ImageNet vs. ProGAN	23.51M	15.27	1.40	4.93	15.50	0.97	19.77
DE-FAKE	MS COCO vs. SD2	308.02M	9.97	29.70	60.40	22.23	55.87	61.00
DIRE	LSUN B. vs. ADM	23.51M	12.53	2.77	1.60	1.93	1.00	20.03
LASTED	LSUN, Danbooru vs. ProGAN, SD1.5	625.63M	71.89	59.59	59.03	52.63	52.43	62.51
J. Ricker 2022	LSUN B. vs. 5 GANs, 5 DMs	23.51M	30.60	73.90	8.50	34.50	6.90	59.76
QAD	7 face swapping datasets	2.56K	38.30	32.47	31.87	33.80	35.43	46.81
InstructBLIP	-	188.84M	97.23	69.10	5.50	3.17	1.60	40.51
InstructBLIP + LoRA	MS COCO vs. SD2	4.19M	99.53	**99.97**	64.30	53.40	50.87	87.30
AntifakePrompt	MS COCO vs. SD2	**4.86K**	99.90	99.93	96.70	93.00	91.57	92.74
	MS COCO vs. SD2+LaMa	**4.86K**	**100.00**	**99.97**	**97.20**	**97.10**	**93.63**	**93.67**

these datasets offer a diverse collection of images depicting people or objects engaged in everyday scenarios, and each of the images is associated with informative caption ground truth. In our work, we selected 90K images, with shorter sides greater than 224, from MS COCO dataset to train our deepfake detector. Moreover, to assess the generalizability of our method over various real images, we additionally select 3K images from Flickr30k dataset to form a held-out testing dataset, adhering to the same criterion of image size.

Generation tasks and generative models In order to evaluate the generalizability and robustness of our model to fake images from emerging and unseen generators, our testing datasets include fake images from 16 different generative models / datasets and 3 distinct attack scenarios , and each of the testing datasets comprises 3K images. We can mainly divide these images into six categories, namely text-to-image generation, Image stylization, image inpainting, super resolution, face swap, and image attacks. The detailed explanation of each category can be found at Appendix B.

It is important to note that for the fake images in the training dataset, we only include images generated by SD2 and SD2IP. Empirical evidence demonstrates that AntifakePrompt, trained solely on these two fake datasets and the real images from MS COCO dataset, exhibits excellent performance on all the other datasets generated by held-out generative models.

Baseline We compare AntifakePrompt to eight recent baseline models, **Wang-2020** (Wang et al., 2020), **DE-FAKE** (Sha et al., 2022), **DIRE** (Wang et al., 2023), **LASTED** (Wu et al., 2023), **J. Ricker 2022** (Ricker et al., 2022), **QAD** (Le & Woo, 2023), **InstructBLIP** (Dai et al., 2023) and **InstructBLIP with LoRA tuning** (Hu et al., 2021). For the detail explanation of the checkpoints we use for every baseline model, please refer to Appendix C.

4.2 Comparisons

AntifakePrompt vs. InstructBLIP without prompt tuning. As depicted in the third and forth rows of Table 1, our detector, trained only on images generated by SD2 and those from MS COCO dataset, exhibits excellent (>85%) performance on most held-in and held-out datasets. In contrast, InstructBLIP without prompt tuning demonstrates generally lower accuracies on most of the testing datasets except for MS COCO, Flickr30k, and LTE. This implies that with the help of prompt tuning, our detector can better understand the deepfake detection task. Thus, our detector is capable of collecting more useful visual features from the input image, resulting in making more accurate decisions on distinguishing real images from fake ones.

AntifakePrompt vs. Baselines. As shown in the first row of Table 1, in contrast to our detector, **Wang 2020** (Wang et al., 2020), trained on images generated by ProGAN (Karras et al., 2017b) and ImageNet (Deng et al., 2009), exhibits satisfactory performance on StyleGAN-XL and yields excellent accuracies on MS COCO and Flickr30k. However, we observe notable decreases in accuracy when it is tested on other held-out datasets. Since they consist of images generated by non-GAN-based models and these images do not share the same artifacts as those in ProGAN-generated images, the detector proposed by (Wang et al., 2020) is unable to differentiate such images by the traits learned from ProGAN-generated images. Regarding **DE-FAKE** (Sha et al., 2022), trained on images generated by SD and MS COCO, it demonstrates impressive performance on real images and 3 diffusion-based models (i.e. SD2, SDXL, and IF) and Deeperforensics, as shown in the second row of Table 1 . However, it struggles to achieve accuracies above 70% on other heldout datasets. Because the detector proposed in (Sha et al., 2022) uses a similar backbone as that in (Wang et al., 2020), it suffers from similar accuracy drops when applying to images generated by unseen generative models. Even though it takes the corresponding prompts into consideration, which allows it to detect unusual scenarios of fake images, it still fails to improve its performance on held-out datasets, since most of them are generated by natural prompts from MS COCO dataset.

As for **DIRE** (Wang et al., 2023), the results are not as excllent as they demostrated in their paper. The possible reason is that our testing dataset (e.g. SD2) comprises images generated by more advanced models than ADM, implying that the distribution of these images is closer to that of real images. Thus, the reconstruction errors of such images are smaller than those of the images generated by ADM, making DIRE harder to differentiate them from real images. Regarding **LASTED** (Wu et al., 2023), we observe performance drops on almost every datasets except for SGXL and 3 face swapping datasets. Since they all employ GAN-based models or models with encoder/decoder structure during their generating process, LASTED can demonstrate relatively high accuracies due to the learned GAN-related artifacts from its training sets. For **J. Ricker 2022** (Ricker et al., 2022), which use the same backbone as that used in (Wang et al., 2020) but trained on more datasets, demonstrates acceptable or even excellent performance on some of the testing datasets, namely diffusion- or GAN-generated datasets. However, similar to what happened to (Wang et al., 2020), it fails to maintain its performance on images generated by unseen models. Lastly, although **QAD** (Le & Woo, 2023) indeed exhibits its excellent performance on 3 face swapping datasets, we observe that it shows relatively low accuracies when testing on other datasets. This indicates that QAD, trained only on 7 face swapping datasets, might not be able to generalize its detection ability to other types of fake images.

Therefore, we can conclude that methods using different strategies or frameworks other than VLM generally demonstrate relatively low accuracies on almost every datasets comprising images generated by unseen models, implying their lacks of generalizability. In contrast, AntifakePrompt can maintain its excellent performance on images generated by unseen models. To discuss the reason, the notable generalizability of LLM, brought by its large training corpus, gives the strong zero-shot ability of VLM, and thus enables AntifakePrompt to show its outstanding generalizability on unseen data.

Also, AntifakePropmt consistently outperforms the other 6 datasets on attacked datasets generated by 3 different attacking strategies. We conclude that our model is more sensitive to the slight and malicious pixel perturbations than its opponents.

Additionally, to address the relatively lower performance observed on LaMa testing dataset, we conduct an experiment to include additional images generated by LaMa into our training dataset. Under this modified setting, as depicted in the forth and fifth rows of Table 1, our detector gives generally

Table 2: **Ablation study: Position of S_* in the tuned prompt (i.e. "prefix", "postfix", or "replace").** The accuracies of real and fake out-of-domain datasets are highlighted in green and red, respectively, and average accuracies are highlighted in blue. The best performances are in **bold**.

Methods	Variant	MS COCO	Flickr	SD2	SDXL	IF	DALLE-2	SGXL	ControlNet	Deeper-Forensics
AntifakePrompt	Replace	95.13	86.20	95.80	93.60	87.33	99.17	98.37	**99.70**	93.03
	Prefix	**95.80**	89.57	97.27	96.47	88.77	97.90	99.87	89.43	94.17
	Postfix	95.37	**91.00**	**97.83**	**97.27**	**89.73**	**99.57**	**99.97**	91.47	**97.90**

Methods	Variant	Inpainting		Super Res.		Attack			Average
		LaMa	SD2	LTE	SD2	Adver.	Backdoor	Data Poisoning	
AntifakePrompt	Replace	33.40	78.63	**99.97**	99.70	90.43	86.00	86.63	88.94
	Prefix	**40.33**	84.67	**99.97**	99.87	93.53	**93.13**	87.43	90.51
	Postfix	39.03	**85.20**	99.90	**99.93**	**96.70**	93.00	**91.57**	**91.59**

Table 3: **Ablation study: Prompt tuning for Q-former, LLM or both.** The accuracies of real and fake out-of-domain datasets are highlighted in green and red, respectively, and average accuracies are highlighted in blue. The best performances are in **bold**.

Methods	Prompt tuning for	MS COCO	Flickr	SD2	SDXL	IF	DALLE-2	SGXL	ControlNet	Deeper-Forensics
AntifakePrompt	Only Q-former	93.50	**92.27**	**97.93**	**98.17**	88.47	99.53	94.80	89.33	**100.00**
	Only LLM	95.10	85.57	95.77	91.73	85.23	98.73	97.80	84.90	93.30
	Both	**95.37**	91.00	97.83	97.27	**89.73**	**99.57**	**99.97**	**91.47**	97.90

Methods	Prompt tuning for	Inpainting		Super Res.		Attack			Average
		LaMa	SD2	LTE	SD2	Adver.	Backdoor	Data Poisoning	
AntifakePrompt	Only Q-former	37.67	77.50	**99.93**	99.67	**97.53**	**97.97**	**95.23**	91.22
	Only LLM	31.37	77.77	99.87	99.40	86.03	83.47	83.50	86.85
	Both	**39.03**	**85.20**	99.90	**99.93**	96.70	93.00	91.57	**91.59**

comparable or even higher accuracies on almost every fake dataset compared to the original setting (the detector trained on 150K training dataset). However, these accuracy enhancements come at the cost of decreased accuracies on real datasets since our detector must now generalize to the inclusion of additional LaMa images in our training set.

4.3 Ablation Study

Position of S_* in the tuned prompt. Comparing the accuracies between InstructBLIP and AntifakePrompt, we empirically show how we ask questions can drastically affect the results. Here, we further investigate how the positioning of pseudo-word S_* in the the tuned prompt can influence our detector. Specifically, we compare 3 different positions of placing pseudo-word: replacing the word ”real” in the prompt with pseudo-word, positioning the pseudo-word in the beginning of the prompt, or placing it at the end of the prompt. For simplicity, we refer to them as ”replace”, ”prefix” and ”postfix”, respectively. As presented in Table 2, although they all yield overall high accuracies, the ”postfix” position exhibits a slight advantage over the other alternatives. This suggests placing the pseudoword at the end of the prompt makes the best efforts among 3 different positions to enable deepfake detection, although the performance difference is not sensitive.

Prompt tuning for the Q-former, LLM or both? We extend our study to the impact of prompt tuning by comparing the results of applying prompt tuning exclusively to Q-former, LLM or both modules. In Table 3, we observe that prompt tuning for both Q-former and LLM outperforms other two alternatives in average accuracy. This implies that prompt tuning for both modules are benefical: tuned prompts to Q-former allow it to extract visual features from input image embeddings that are more conducive to differentiating between real and fake images. Tuned prompts to LLM can more precisely describe the idea of fake image detection for LLM, and thus, LLM is able to make more accurate decisions on differentiating real and fake images. Due to the improved visual features and improved instruction, the application of prompt tuning for both Q-former and LLM yields better performance.

Number of training images. We first study the effect of the number of real images in the training dataset. While fixing the number of fake images in training dataset to 60K, we gradually increase the number of real images from 30K to 120K in the step of 30K, resulting in the total size of our training dataset ranging from 90K to 180K. As shown in Row 1 to 4 in Table 4, while the accuracies

Table 4: **Ablation study: Number of training datasets,** where either the number of the real images or that of both real and fake images in the training dataset is reduced. The accuracies of real and fake out-of-domain datasets are highlighted in green and red, respectively, and average accuracies are highlighted in blue. The best performances are in **bold**.

Methods	No. Data	MS COCO	Flickr	SD2	SDXL	IF	DALLE-2	SGXL	ControlNet	Deeper-Forensics
AntifakePrompt	90K	89.90	80.37	**98.33**	**98.20**	**92.87**	98.90	99.93	**94.60**	**99.27**
	120K	94.30	89.10	97.53	96.60	89.20	98.27	99.90	91.03	95.13
	150K	95.37	91.00	97.83	97.27	89.73	99.57	**99.97**	91.47	97.90
	180K	96.53	92.23	97.37	96.67	88.43	**99.63**	99.93	90.43	94.63
	15K	92.60	92.03	92.17	90.20	75.80	97.27	97.97	81.27	91.93
	1.5K	92.83	94.10	75.67	68.20	58.17	79.77	90.63	67.50	67.53
	0.15K	**99.07**	**99.73**	33.17	16.73	16.30	36.93	56.73	26.17	24.20

Methods	No.	Inpainting		Super Res.			Attack		Average
	Data	LaMa	SD2	LTE	SD2	Adver.	Backdoor	Data Poisoning	
AntifakePrompt	90K	**49.80**	**89.63**	**100.00**	**99.93**	**98.40**	**96.40**	94.97	**92.59**
	120K	41.03	85.13	99.97	99.87	97.87	96.00	**95.37**	91.64
	150K	39.03	85.20	99.90	**99.93**	96.70	93.00	91.57	91.59
	180K	36.87	84.27	99.73	**99.93**	94.83	91.50	90.57	90.85
	15K	34.73	73.93	**100.00**	98.43	84.07	82.57	74.53	84.97
	1.5K	30.80	67.57	99.87	90.00	29.57	25.93	11.87	65.63
	0.15K	8.90	40.53	95.23	63.53	3.97	2.40	1.43	39.06

on testing datasets of real images increase along with the increments of real images in training dataset, our detector suffers from a decrease in accuracies on fake testing datasets. Therefore, we designate the detector trained on 150K training dataset as our optimal model, achieving balanced accuracies on real and fake images.

To explore the limit of few-shot learning ability of our detector, we gradually reduce the numbers of both real images and fake images in training dataset to one tenth at each step until only 150 images remain in total. As shown in Row 5 to 7 in Table 4, we found out that our detector still outperforms DE-FAKE in almost every testing datasets (except for SD2, SDXL and IF) when there are as few as 15K training samples, almost a quarter of DE-FAKE. Additionally, when we reduce our training dataset to only 1.5K images, only 0.2% of (Wang et al., 2020) training dataset, our detector outperforms (Wang et al., 2020) on every fake dataset, exhibiting only slightly lower accuracies on real datasets. These findings underscore the data effiency of our detector in terms of training data size compared to eight baseline models.

Finetuning InstructBLIP with LoRA. Furthermore, we conduct extended experiments to compare between our prompt tuning and LoRA-based (Hu et al., 2021) InstructBLIP parameter finetuning. The results, as shown Row 4 and 5 in Table 1 corresponding to "AntifakePrompt" with training set "MS COCO vs. SD2" for our model performance to make comparison, reveal that while the detector finetuned with LoRA achieves comparable results in certain testing datasets, our detector consistently outperforms it in the three attack datasets. This underscores the sensitivity of our detector to such attack scenarios. Since additional LoRA matrices introduce relatively more learnable parameters into LLM (around 4M) than those introduced by prompt tuning (around 4K), it is more likely for LoRA-tuned InstructBLIP to overfit to artifacts of training datasets, resulting in accuracy drops when applied to fake datasets with different traits, namely three attack datasets.

5 Conclusion

In this paper, we propose a solution to deepfake detection problem utilizing vision-language model to address the limitations of traditional deepfake detection methods when being applied on held-out dataset. We formulate the deepfake detection problem as a visual question answering problem, and apply soft prompt tuning on InstructBLIP. Empirical results demonstrate improved performance of our detector over both held-in and held-out testing datasets, which is trained solely on generated images using Stable Diffusion 2 and real images from MS COCO datasets. Furthermore, in contrast to prior studies which require to finetune/learn millions of parameters, our model only needs to tune 4864 trainable parameters, thus striking a better balance between the training cost and the effectiveness. Consequently, our detector provides a potent defense against the potential risks associated with the misuse of generative models, all while demanding fewer training resources.

Reproducibility Statement

In accordance with the principles of reproducibility and to foster further research explorations, we provide open access (Appendix A) to all resources related to our study as part of the supplementary materials, including a complete and documented codebase with all models, scripts, configs, checkpoints, preprocessing, training and evaluation codes that can reproduce the results shown in the paper. Furthermore, we promise to maintain these resources and offer the necessary support for any clarification or query associated with the public available resources.

References

Stable-diffusion-v1-5. https://huggingface.co/runwayml/stable-diffusion-v1-5.

Danbooru: A large-scale crowd sourced and tagged anime illustration dataset, 2021. https://www.gwern.net/Danbooru2021.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Lichang Chen, Jiuhai Chen, Tom Goldstein, Heng Huang, and Tianyi Zhou. Instructzero: Efficient instruction optimization for black-box large language models. *arXiv preprint arXiv:2306.03082*, 2023.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8628–8638, 2021.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

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.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric P Xing, and Zhiting Hu. Rlprompt: Optimizing discrete text prompts with reinforcement learning. *arXiv preprint arXiv:2205.12548*, 2022.

Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset, 2019.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Jonas Geiping, Liam Fowl, W Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. *arXiv preprint arXiv:2009.02276*, 2020.

Oliver Giudice, Luca Guarnera, and Sebastiano Battiato. Fighting deepfakes by detecting gan dct anomalies. *Journal of Imaging*, 7(8):128, 2021.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Level up the deepfake detection: a method to effectively discriminate images generated by gan architectures and diffusion models. *arXiv preprint arXiv:2303.00608*, 2023.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yang He, Ning Yu, Margret Keuper, and Mario Fritz. Beyond the spectrum: Detecting deepfakes via re-synthesis. *arXiv preprint arXiv:2105.14376*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Ivan Itzcovich. Faced, 2018. `https://github.com/iitzco/faced`.

Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2889–2898, 2020.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017a.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017b.

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*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021.

Hoki Kim. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*, 2020.

Binh M Le and Simon S Woo. Quality-agnostic deepfake detection with intra-model collaborative learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22378–22389, 2023.

Jaewon Lee and Kyong Hwan Jin. Local texture estimator for implicit representation function. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1929–1938, 2022.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.

Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 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.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16463–16472, 2021.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 725–741. Springer, 2020.

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

Peter Lorenz, Ricard L Durall, and Janis Keuper. Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 448–459, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ruipeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. Exposing the fake: Effective diffusion-generated images detection. *arXiv preprint arXiv:2307.06272*, 2023.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 3195–3204, 2019.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 947–952. IEEE, 2019.

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.

Alex 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. *arXiv preprint arXiv:2112.10741*, 2021.

OpenAI. GPT-4 Technical Report. *arXiv e-prints*, art. arXiv:2303.08774, March 2023. doi: 10.48550/arXiv.2303.08774.

OpenAI. Dalle-3, 2023. https://openai.com/dall-e-3.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

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.

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*, pp. 1–11, 2019.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022a.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. volume abs/2201.00273, 2022b. URL https://arxiv.org/abs/2201.00273.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162. Springer, 2022.

Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image diffusion models. *arXiv preprint arXiv:2210.06998*, 2022.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

StabilityAI. Deepfloyd if, 2023. https://github.com/deep-floyd/IF.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2149–2159, 2022.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2387–2395, 2016.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8695–8704, 2020.

Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. *arXiv preprint arXiv:2206.02262*, 2022.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. *arXiv preprint arXiv:2303.09295*, 2023.

Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *arXiv preprint arXiv:2302.03668*, 2023.

Haiwei Wu, Jiantao Zhou, and Shile Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv preprint arXiv:2305.13800*, 2023.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7556–7566, 2019.

Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pp. 7354–7363. PMLR, 2019a.

Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

Tianjun Zhang, Xuezhi Wang, Denny Zhou, Dale Schuurmans, and Joseph E Gonzalez. Tempera: Test-time prompt editing via reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2022.

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)*, pp. 1–6. IEEE, 2019b.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Andy Zou, Zifan Wang, J Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023.

A Links to resources

- Link to codes, chechpoints of AntifakePrompt (anonymous GitHub): `https://anonymous.4open.science/r/LAVIS-C8E1/README.md`
- Link to checkpoints of AntifakePromt (Google Drive): `https://drive.google.com/drive/folders/1JgMJie4wDt7dNeHkT25VVuzG9CdnA9mQ?usp=drive_link`

B Generation of testing datasets

1. **Text-to-images generation**: We collected 3K prompts, half of which are sampled from the MS COCO ground truth captions and the other half from Flickr30k captions. These prompts are then input into five different generative models, i.e. SD2 (Rombach et al., 2022), SDXL (Podell et al., 2023), DeepFloyd IF (StabilityAI, 2023), DALLE-2 (Ramesh et al., 2022), SGXL (Sauer et al., 2022a) , and GLIDE (Nichol et al., 2021), to generate the corresponding images.
2. **Image stylization**: We begin by extracting Canny edge features from the 3000 test images in MS COCO dataset mentioned in the previous paragraph. Subsequently, we pass these Canny edge feature images, along with the corresponding prompts, into ControlNet (Zhang & Agrawala, 2023) to generate stylized images.
3. **Image inpainting**: We employ the same 3000 test images and resize them to make the shorter side of each image be 224, which matches the input size of InstructBLIP. Then, we randomly generate masks of three distinct thickness levels for these resized images using the scripts from the LaMa (Suvorov et al., 2022) GitHub[5]. With original images and the corresponding masks prepared, we utilize two different models, SD2-Inpainting (SD2IP) and LaMa, to inpaint images, respectively. The resizing step ensures that most of artifacts created during the inpainting process will be retained before being inputted to the detector.
4. **Super Resolution**: Out of the same reason in the inpainting, we apply the same resizing process to the same 3000 test images before downsizing them to one-forth of their original size. These low-resolution images are then passed into two different models, SD2-SuperResolution (SD2SR) and LTE (Lee & Jin, 2022), to upsize back. A scaling factor of four is chosen, as only the ×4-upscaling weights for SD2 are publicly available.
5. **Face Swap**: Since face swapping is also one of the common means to generate fake images, we employ three large-scale face swapping video datasets, namely Deeperforensics (Jiang et al., 2020), DFDC (Dolhansky et al., 2019) and FaceForensics++ (Rossler et al., 2019). From each of these datasets, we randomly extract frames from 1000 randomly selected videos. Following (Wang et al., 2020), we then apply Faced (Itzcovich, 2018) to crop out 3000 faces from the extracted frames of each dataset to ensure that complete facial features are present in every image.
6. **Image attacks**: We apply three common types of attacks to edit images and target at a traditional ResNet-50 classifier. The attack types include adversarial attack (Kim, 2020), backdoor attack (Li et al., 2021) and data poisoning attack (Geiping et al., 2020). Default settings are employed for each attack. By testing our detector on these attacks, we can have a better understanding of its sensitivity against these slight and malicious image editing.

C Checkpoints of each baseline

This section lists the checkpoint details of all the baseline mentioned in 1.

1. **Wang-2020**: We use the detector checkpoint that is trained on dataset with images that are possibly Gaussian blurr- and JPEG-augmented, each with 10% probability.
2. **DE-FAKE**: We use the checkpoint of the hybrid detector, which considers both the image and the corresponding prompts during detection.

[5] `https://github.com/advimman/lama/blob/main/bin/gen_mask_dataset.py`

3. **DIRE**: We use the checkpoint of detector trained on images from LSUN-Bedroom (LSUN B.) (Yu et al., 2015) and those generated by ADM.
4. **LASTED**: We use the checkpoint of the detector trained on images from LSUN (Yu et al., 2015) and Danbooru (Dan, 2021), and those generated by ProGAN (Karras et al., 2017a) and SD1.5 (SD1).
5. **J. Ricker 2022**: We use the checkpoint that trained on images generated by 5 GANs and 5 DMs (i.e. ProGAN, StyleGAN, ProjectedGAN (Sauer et al., 2021), Diff-StyleGAN2 (Wang et al., 2022), Diff-ProjectedGAN (Wang et al., 2022), DDPM, IDDPM, ADM , PNDM (Liu et al., 2022) and LDM).
6. **QAD**: We use the checkpoint of detector trained on 7 face swapping datasets (i.e. NeuralTextures (Thies et al., 2019), Deepfakes, Face2Face (Thies et al., 2016), FaceSwap (Thies et al., 2016), FaceShifter (Li et al., 2019), CelebDFv2 (Li et al., 2020) and FaceForensicsIntheWild).
7. **InstructBLIP**: We use the pretrained weight provided by LAVIS and preset question prompt without prompt tuning.
8. **InstructBLIP with LoRA**: We also use the pretrained weight provided by LAVIS and the preset question prompt, but apply LoRA tuning on LLM of InstructBLIP instead of prompt tuning.

D Samples for each datasets

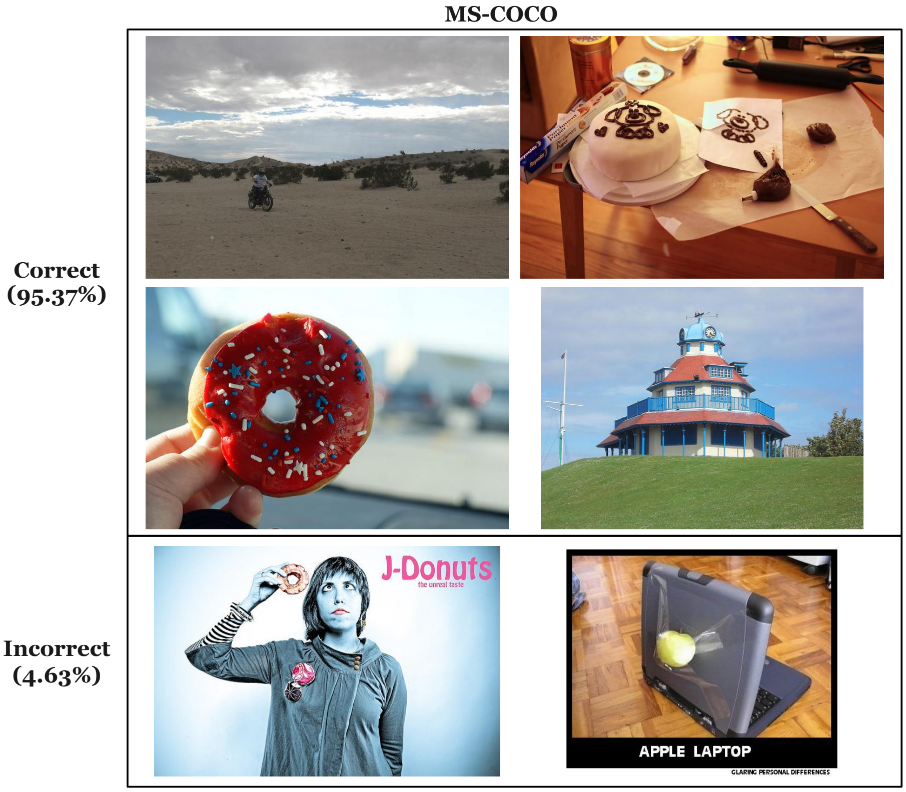

Figure 2: **Samples for each datasets.** 95.37% of images are correctly classified as real.

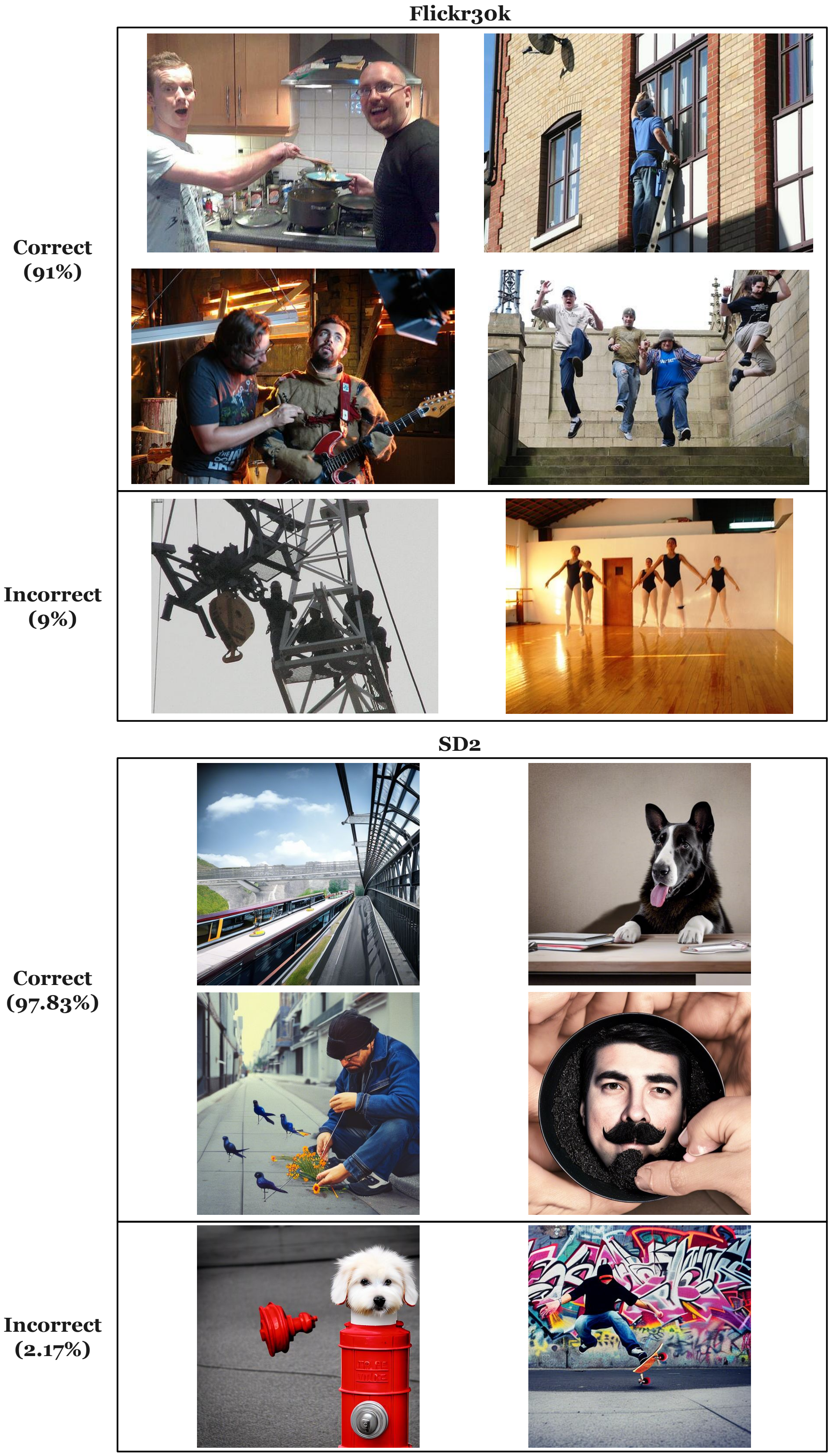

Figure 3: **Samples for each datasets (Continue).** 91% / 97.83% of images in Flickr30k / generated by SD2 are correctly classified as real / fake.

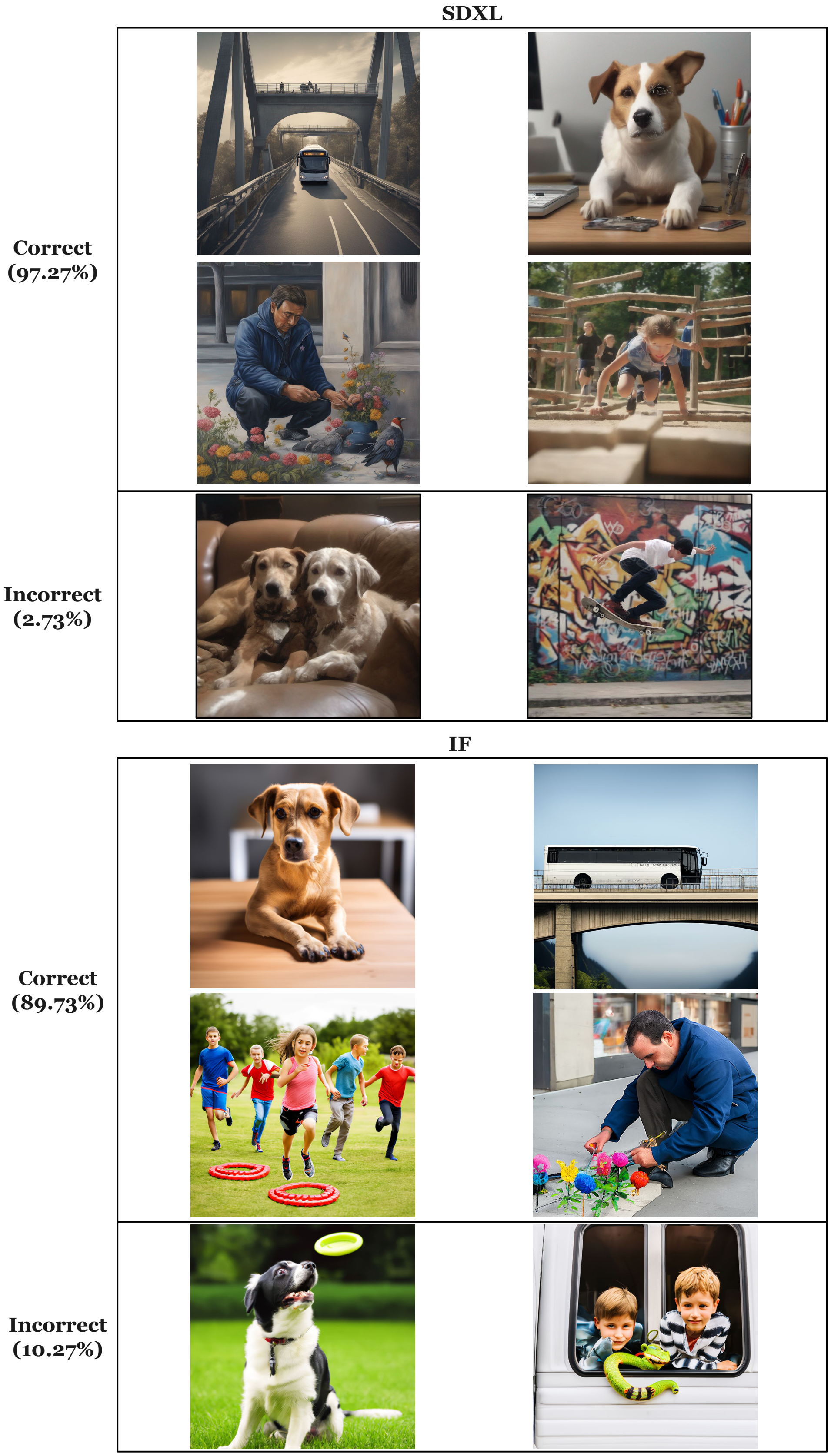

Figure 4: **Samples for each datasets (Continue).** 97.27% / 89.73% of images generated by SDXL / IF are correctly classified as fake.

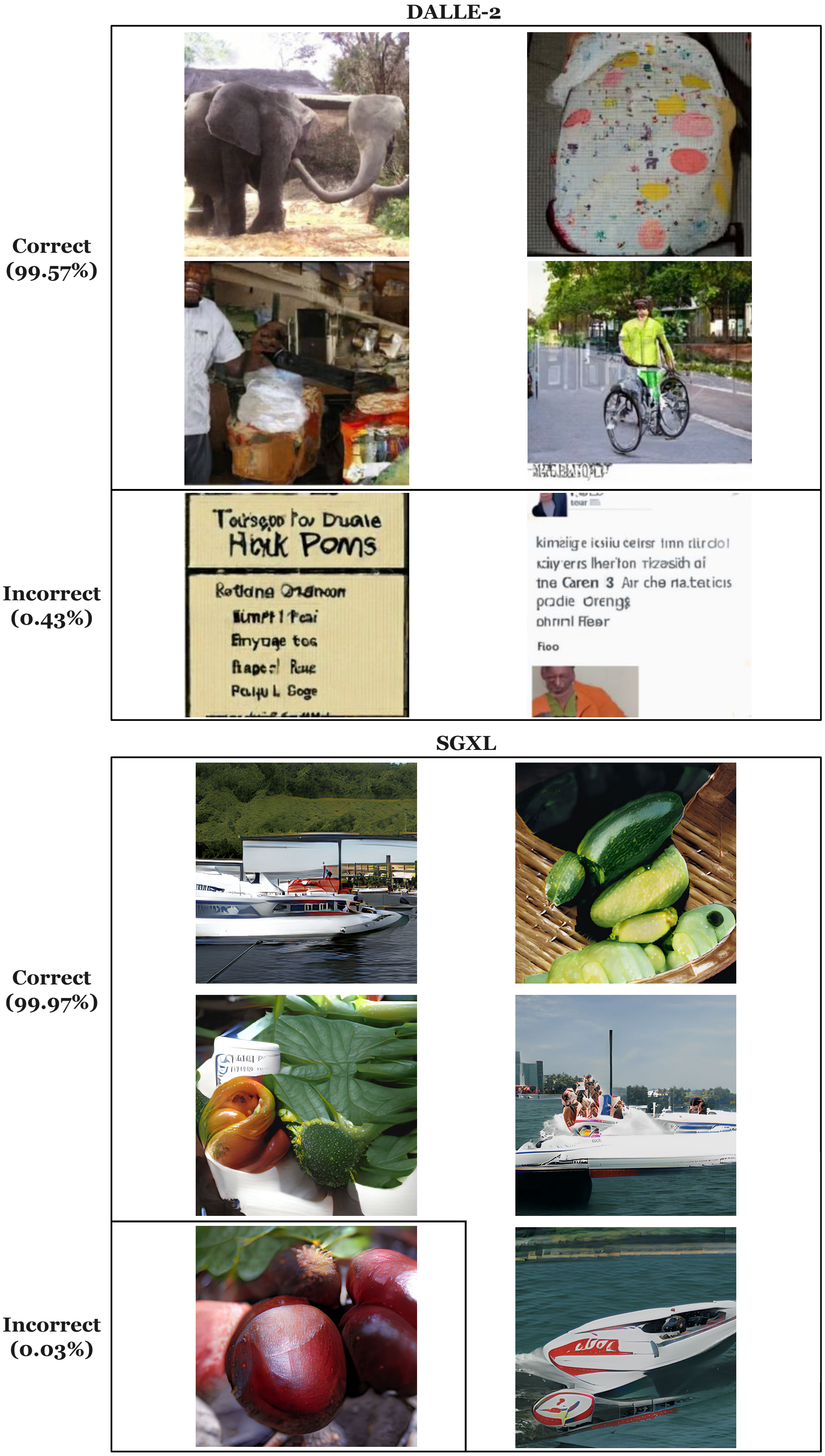

Figure 5: **Samples for each datasets (Continue).** 99.57% / 99.97% of images generated by DALLE-2 / SGXL are correctly classified as fake.

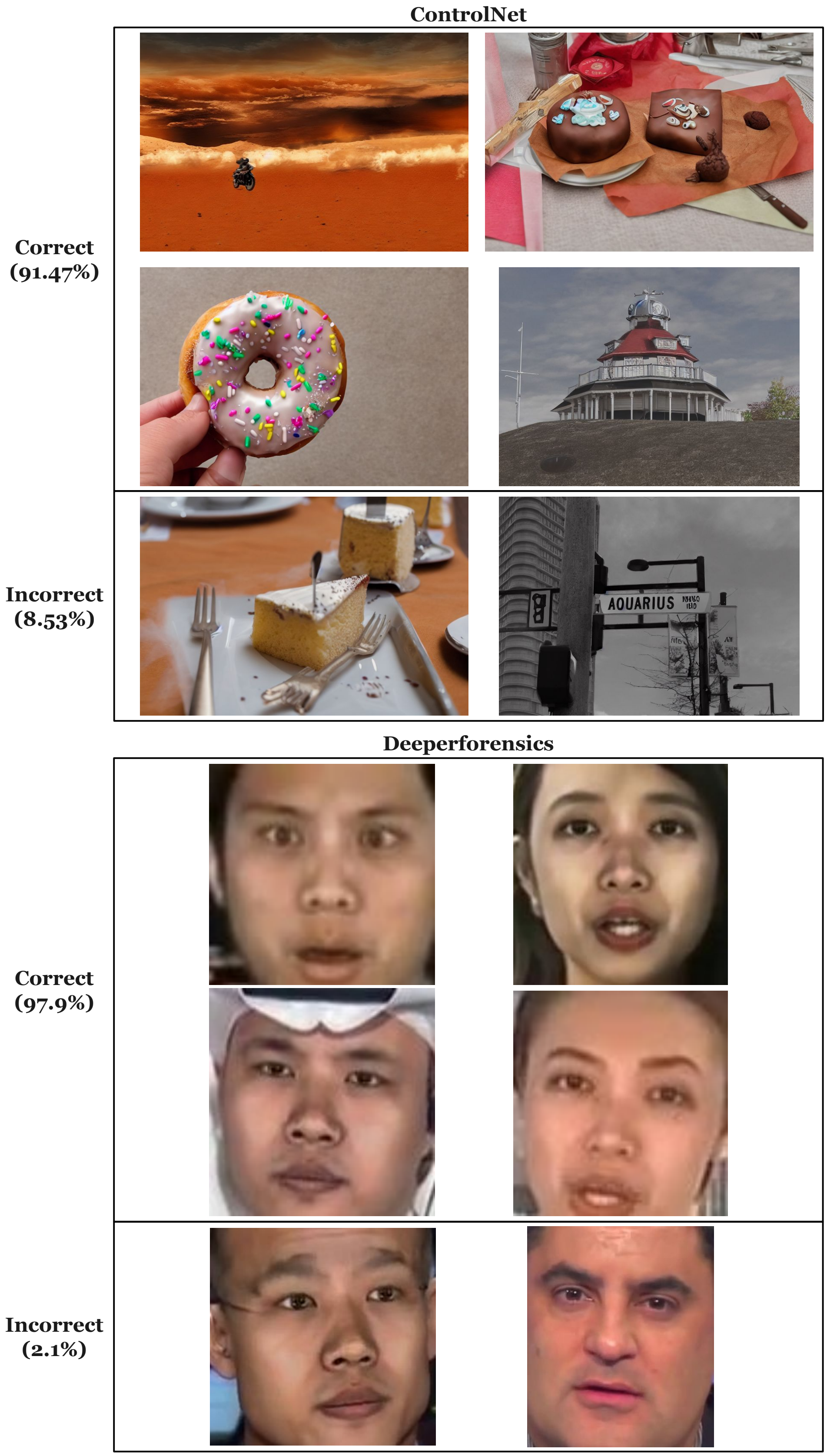

Figure 6: **Samples for each datasets (Continue).** 91.47% / 97.9% of images generated by ControlNet / in Deeperforensics are correctly classified as fake.

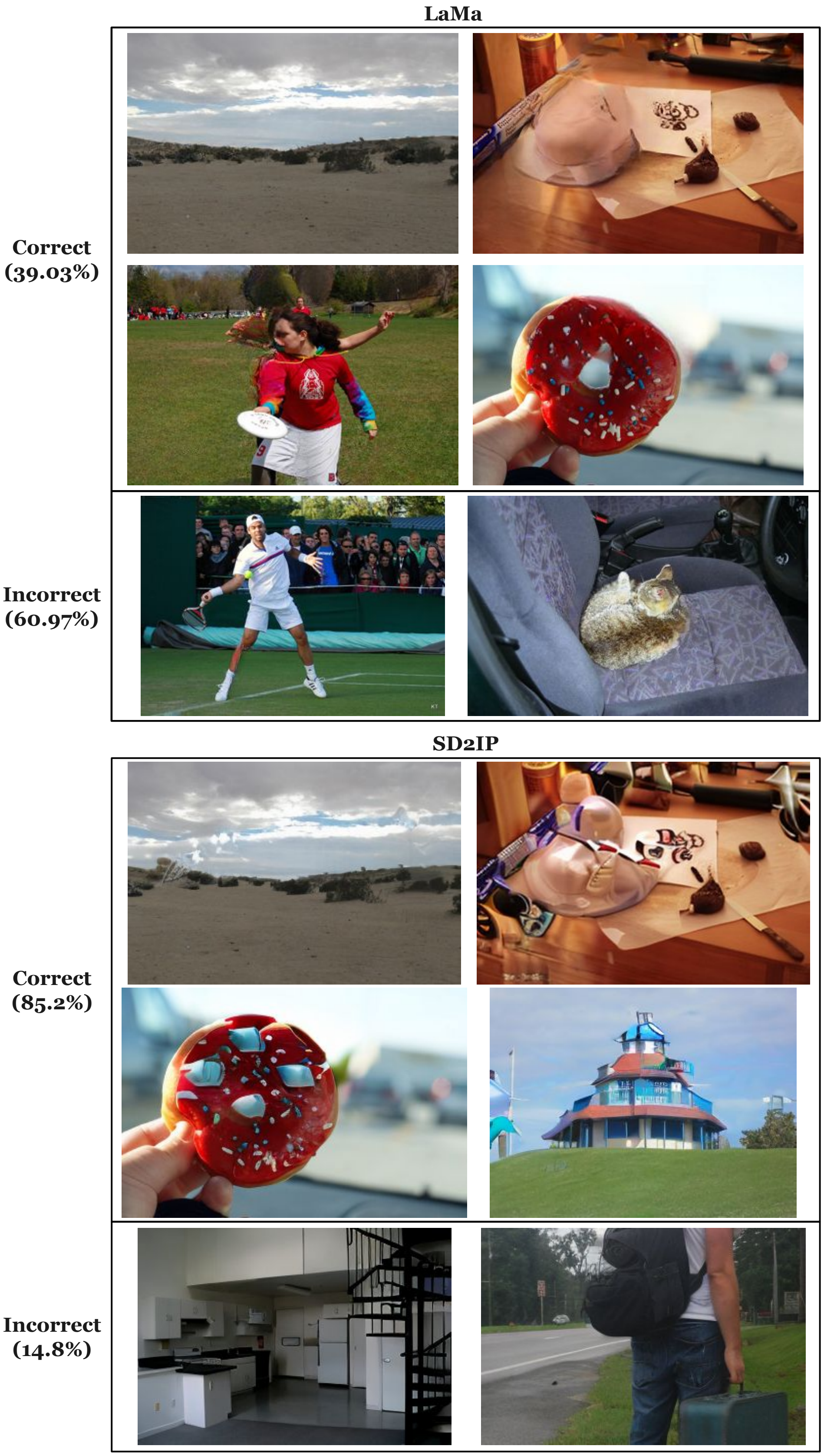

Figure 7: **Samples for each datasets (Continue).** 39.03% / 85.2% of images generated by LaMa / SD2IP are correctly classified as fake.

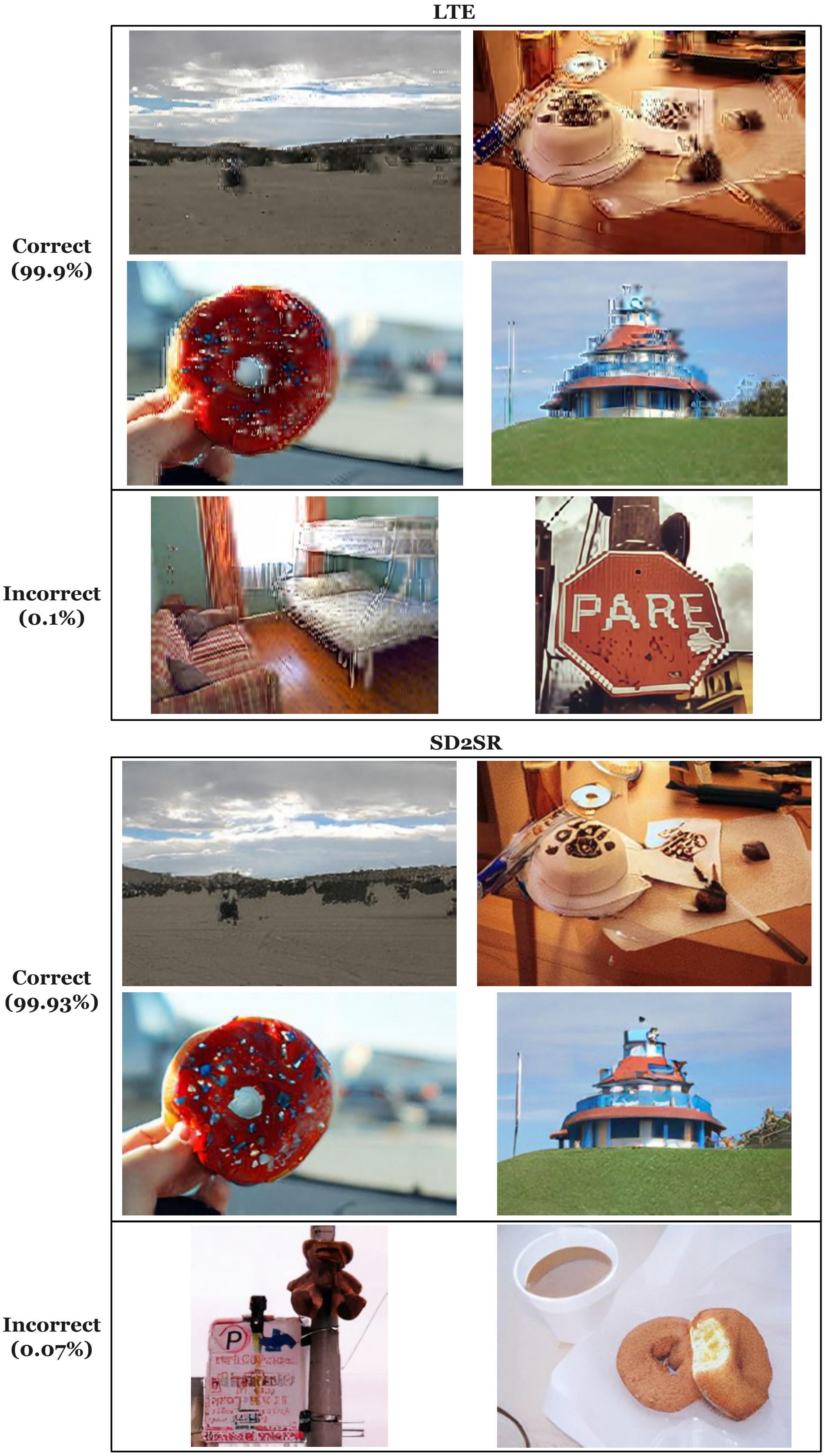

Figure 8: **Samples for each datasets (Continue).** 99.9% / 99.93% of images generated by LTE / SD2ISR are correctly classified as fake.

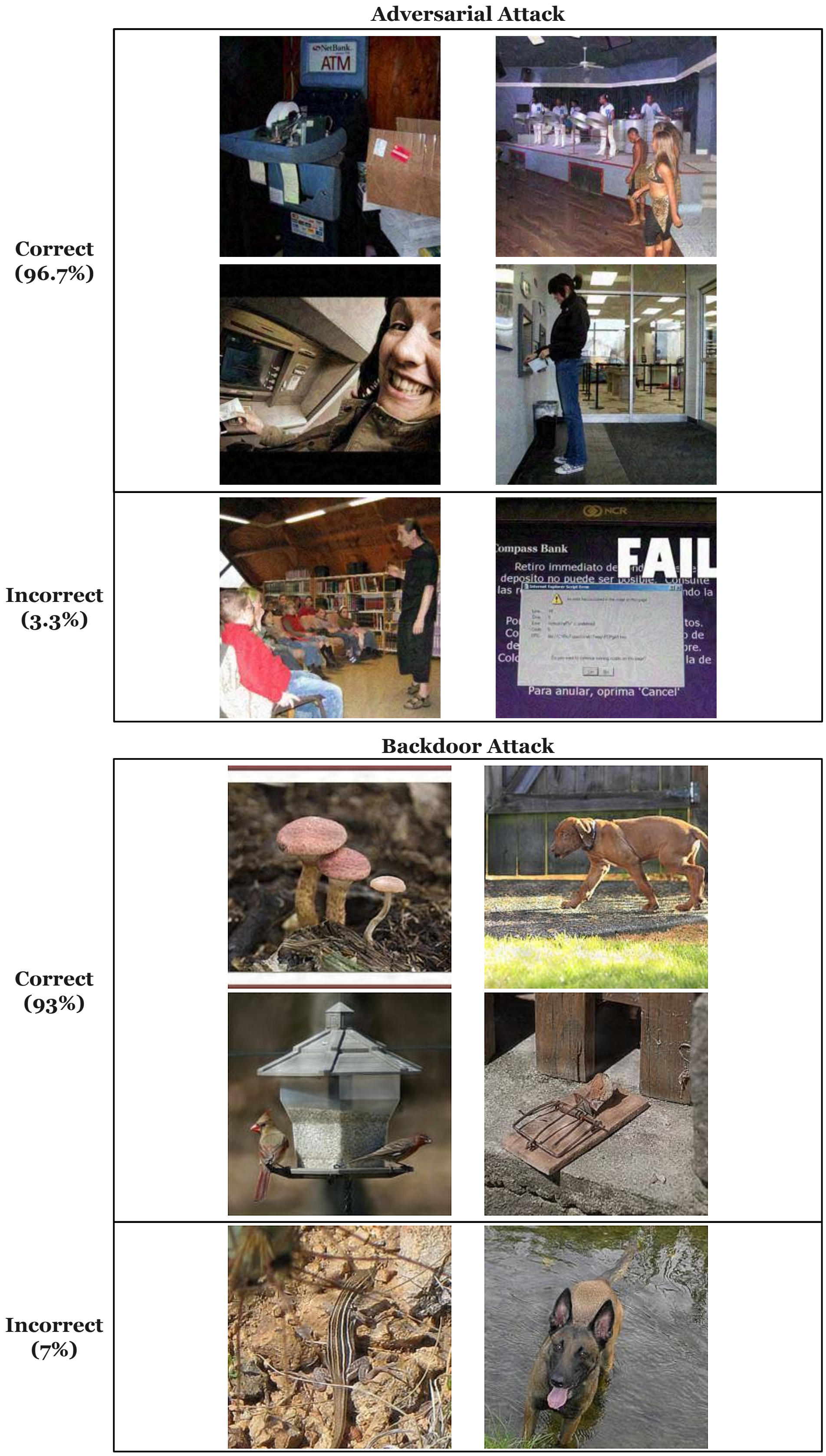

Figure 9: **Samples for each datasets (Continue).** 96.7% / 93% of images generated under adversarial / backdoor attack are correctly classified as fake.

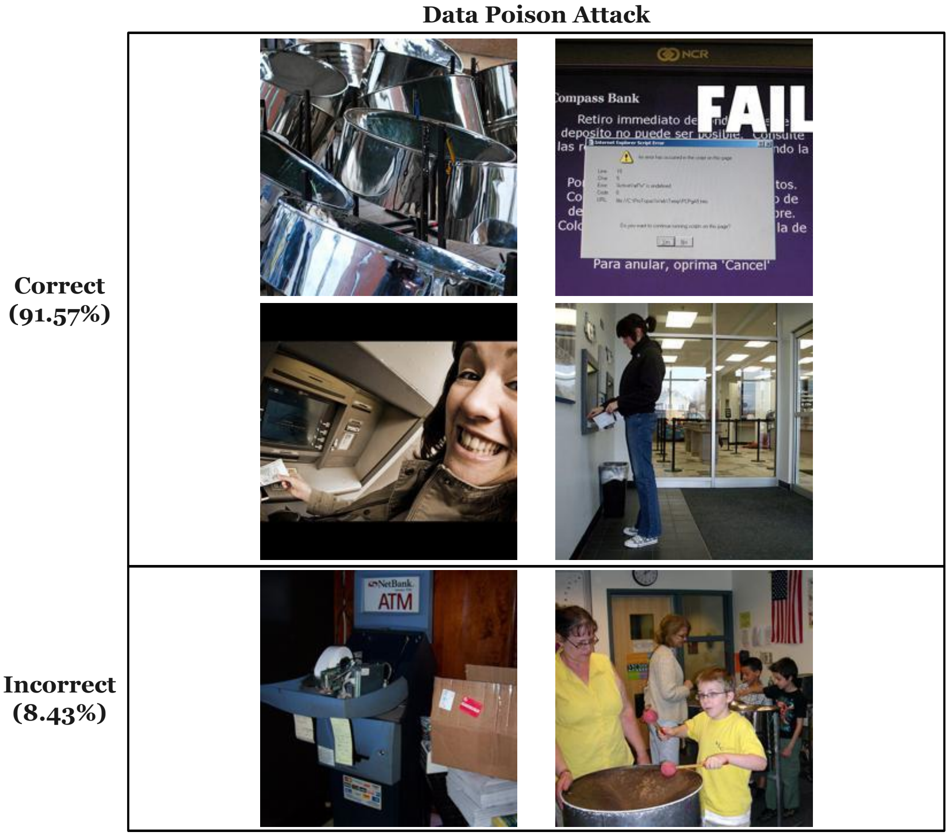

Figure 10: **Samples for each datasets (Continue).** 91.57% of images generated under data poisoning attack are correctly classified as fake.