# Visual Language Models as Zero-Shot Deepfake Detectors

**Viacheslav Pirogov** [1]

## Abstract

The contemporary phenomenon of deepfakes, utilising GAN or diffusion models for face swapping, presents a substantial and evolving threat in digital media, identity verification, and a multitude of other systems. The majority of existing methods for detecting deepfakes rely on training specialised classifiers to distinguish between genuine and manipulated images, focusing only on the image domain without incorporating any auxiliary tasks that could enhance robustness.

In this paper, inspired by the zero-shot capabilities of Vision-Language Models (VLMs), we propose a novel approach using VLMs to identify deepfakes. We introduce a new, high-quality deepfake dataset comprising 60,000 images, on which zero-shot VLMs demonstrate superior performance to almost all existing methods. Subsequently, we compare the performance of the best-performing VLM, InstructBLIP, on the popular deepfake dataset DFDC-P against traditional methods in two scenarios: zero-shot and in-domain fine-tuning. Our results demonstrate the superiority of VLMs over traditional classifiers.

## 1. Introduction

The recent advancements in generative computer vision have rendered numerous technologies that were previously deemed unfeasible accessible to the general public. The advent of open computing tools such as Google Colab (Google LLC) has democratized access to high-quality and high-fidelity face generation techniques (Wu et al., 2023a; Liu et al., 2021; Gecer et al., 2021), and precise manipulation of facial attributes (Mou et al., 2023; Kim et al., 2022; Hou et al., 2022) using publicly available open-source code. While the democratisation of these technologies has numerous positive aspects, it also facilitates the spread of fake news (Zhou et al., 2020; Huang et al., 2023a), impersonation (Li et al., 2020a), false authentication, and liveness bypassing (Sabaghi et al., 2021). In this context, deepfakes — highly realistic digital manipulations of human faces — have emerged as a powerful tool for the dissemination of misinformation and the perpetration of identity theft.

Many works have emphasized that deepfake detection remains a significant unsolved problem in the modern world (Le et al., 2024; Liu et al., 2024b; Heidari et al., 2024; Le et al., 2023). One of the primary challenges is the lack of a comprehensive dataset that encompasses all types of deepfakes encountered in real-life scenarios. Furthermore, the existing detectors lack robustness against simple manipulations such as noise (Haliassos et al., 2021; Jiang et al., 2020) or compression (Le & Woo, 2023; 2021). Although detectors can achieve high accuracy on tested datasets, they frequently fail to generalise to new types of deepfakes. It would be optimal to have a robust zero-shot or few-shot method that can effectively detect previously unseen deepfakes. ChatGPT (Fraser, 2023) and other large language models (LLMs) (Team, 2024a) have demonstrated excellent zero-shot and generalization capabilities. In light of this, Visual Language Models (VLMs) with instruction tuning (Dai et al., 2023; Liu et al., 2023; Laurençon et al., 2024) have demonstrated comparable potential, making them well-suited for the task of deepfake detection.

Previous studies have demonstrated the potential of both open-source and closed-source Visual Language Models (VLMs) in the task of deepfake detection (Chang et al., 2023; Zhang et al., 2024; Li et al., 2024; Shi et al., 2024; Jia et al., 2024). However, these studies have not fully explored the pure zero-shot capabilities of VLMs and have not focused on integrating such models into real-life systems, such as liveness checks and verifications.

In this work, we propose a novel method to distinguish between real and fake images using Visual Language Models (VLMs). We demonstrate the superiority of this approach by creating a new high-quality deepfake dataset and evaluating state-of-the-art deepfake detectors alongside open-source and closed-source VLMs in zero-shot and few-shot setups on this dataset. Furthermore, we show that through language fine-tuning on the widely used deepfake dataset DFDC-P (Dolhansky et al., 2020), the best-performing VLM is at

[1] Sumsub, Berlin, Germany. Correspondence to: Viacheslav Pirogov <slava.pirogov@sumsub.com>.

*Proceedings of the $42^{nd}$ International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

least as effective as previous approaches.

## 2. Related work

This section provides an overview of the techniques used to generate deepfakes in the modern world, with a particular focus on face swapping between two individuals 2.1 and methods for detecting such deepfakes 2.2. Subsequently, we present the current state of Visual Language Models (VLMs), with a particular focus on the VLMs that we have utilised 2.3. In the last subsection 2.4, we offer a comprehensive overview of existing research that employs VLMs as deepfake detectors.

### 2.1. Deepfake Generation

Face swapping approaches can be categorized based on the number of images required for the source and target faces. These approaches range from using large datasets, commonly referred to as ”facesets”, to employing few-shot or one-shot methods. The most powerful method utilising large datasets is DeepFaceLab (Perov et al., 2020). In contrast, few-shot or one-shot methods (Chen et al., 2020; Nirkin et al., 2019; 2022; Li et al., 2019; Jia Guo, 2017) are the most popular today within the community (Sangwan, 2023a;b; C0untFloyd, 2023; machineminded, 2024) and among fraudsters due to their ease of use and low resource requirements.

**SimSwap**: An Efficient Framework For High Fidelity Face Swapping is a state-of-the-art (SOTA) open-source model for high-fidelity one-shot face swapping, requiring only one image each from source and target. This model supports relatively high resolutions with two variants: one at 224x224 pixels, trained on the VGGFace2 (Cao et al., 2018) dataset, and other at 512x512, trained on enhanced VGGFace2-HQ dataset. SimSwap utilises an adversarially trained encoder-decoder architecture, augmented by an Identity Injection Module, which separates image attributes and identity, thereby enabling the transfer of only the attributes. To this end, a face recognition network (Deng et al., 2019) is employed to extract an identity embedding, which is then integrated via Adaptive Instance Normalization (AdaIN) (Huang & Belongie, 2017). The model is trained in a GAN style (Goodfellow et al., 2014; Liu et al., 2019; Brock et al., 2019; Gulrajani et al., 2017; Karras et al., 2019; Isola et al., 2017) to ensure the realism of generated images, with the objective of achieving an indistinguishable result from that of the original.

### 2.2. Deepfake Detection

For the last five years, researchers have been actively working on deepfake detection methods. They usually propose new datasets that better represent real-life scenarios than previous ones (Rossler et al., 2019; Dolhansky et al., 2020; Li et al., 2020b; Jiang et al., 2020; Shiohara & Yamasaki, 2022) or introduce new analytical approaches such as novel architectures ((Zhao et al., 2021; Wang et al., 2022; Sun et al., 2021b)), frequency-based methods (Le & Woo, 2021; Qian et al., 2020; Song et al., 2022), spatial techniques (Le & Woo, 2023; Nguyen et al., 2018; Tariq et al., 2021), and many other approaches (Cao et al., 2022; Dong et al., 2023; Sun et al., 2024; Chen et al., 2021; Sun et al., 2021a).

The first significant work in deepfake area was **FaceForensics++** (Rossler et al., 2019), which presented a dataset of over 1.8 million images from 1000 YouTube videos, accompanied by a simple detection model based on XceptionNet (Chollet, 2017). Building on this, the authors of **MAT** (Zhao et al., 2021) proposed moving from a simple CNN model to a multi-attention network, inspired by the popularity of Visual Transformers (Dosovitskiy et al., 2020; Vaswani et al., 2017). Similarly, the authors of **M2TR** (Wang et al., 2022) employed a frequency filter (Ricker et al., 2022) with a 2D Fast Fourier Transform to enhance detection. Another innovative approach is **RECCE** (Cao et al., 2022), which employs metric-learning loss and reconstruction learning (Wertheimer et al., 2021) to improve upon previous methods. One of the most effective and robust methods is **SBI** (Shiohara & Yamasaki, 2022), which presents a unique dataset generated by blending pseudo-source and target images derived from individual pristine images.

### 2.3. Visual Language Models

The first significant work in the modern state of Visual Language Models (VLMs) field was Flamingo, introduced in (Alayrac et al., 2022). The authors proposed utilising pretrained and frozen during fine-tuning Vision Encoder and Language Model (LM), connected by a Perceiver Resampler (Jaegle et al., 2021), which processes varying-size large feature maps and outputs few visual tokens. These tokens are then fed through gated cross attention and trained with language modelling loss. Another notable initial work was CoCa (Yu et al., 2022), where the authors combined contrastive pre-training and language modelling into a single model. This was achieved by passing image and text pairs to the corresponding encoders, where a contrastive loss was calculated on the CLS tokens. Subsequently, a text decoder with cross-attention on image features was employed to calculate the language modelling loss.

One of the most significant open-source VLM families is the LAVIS family (Salesforce, 2022). The initial work ”BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation” was introduced in (Li et al., 2022). The authors proposed an encoder-decoder architecture trained on three tasks: contrastive image-text pairing (Radford et al., 2021), image-text matching with cross-attention, and language modeling loss. A crucial part of this work was dataset bootstrapping with synthetic captions. The continuation of this work is BLIP-2 (Li et al., 2023), which aimed to combine a pre-trained and frozen image encoder and LLM with a lightweight module named Q-Former, containing a few self and cross-attention layers. Inspired by zero-shot capabilities of instruction tuning (OpenAI, 2022), the authors presented InstructBLIP (Dai et al., 2023), which involved the collection of a new instruction dataset and the fine-tuning of the Q-Former and frozen LLM, resulting in improved overall performance.

In 2023-2024, VLMs diverged into two main approaches: training large models on extensive datasets with different tasks (Alayrac et al., 2022; Li et al., 2022; Yu et al., 2022), and using a small connector between frozen Vision Encoders and LLMs (Li et al., 2023; Dai et al., 2023; Liu et al., 2023; Laurençon et al., 2024). In 2023, the authors of FROMAGe (Koh et al., 2023) demonstrated that connecting a pre-trained Visual Encoder and LLM could be achieved with just three linear layers as a projection and the addition of a special image token, with training for only one day on a single GPU. This straightforward and cost-effective approach outperformed many previous methods that had been trained on multiple GPUs over extended periods.

Inspired by FROMAGe (Koh et al., 2023) and instruction tuning (OpenAI, 2022), LLaVA proposed in (Liu et al., 2023), with a creation of a new instruction image-language dataset. The authors used ChatGPT (OpenAI, 2022) to generate captions and questions in a chatbot format without seeing the picture. Following 150 hours of training with FROMAGe-like architecture, the model became SOTA in many benchmarks, even surpassing closed GPT-4V (Team, 2024b). Subsequently, the authors released LLaVA 1.5 and 1.6 (LLaVA-NeXT) (Liu et al., 2024a), with minor improvements in training data and trained models using other pre-trained LLMs. One of the most recent models developed by Hugging Face, called Idefics2 (Laurençon et al., 2024), is very similar to LLaVA, with Modality Projection and Pooling as connectors and its own instruction dataset.

### 2.4. VLMs in deepfake detection

Researchers have already shown some potential of Visual Language Models (VLMs) in the deepfake detection task. Nevertheless, the full generalisability of these models in zero-shot or few-shot setups has not yet been demonstrated. In this subsection, we review existing methods, some of which focus on fine-tuning (Chang et al., 2023), while others reformulate the classification task into reasoning or Visual Question Answering (VQA) tasks (Zhang et al., 2024; Li et al., 2024). A few studies also explore how to employ closed-source VLMs such as GPT-4V (Team, 2024b) and Gemini (Team, 2024a) (Shi et al., 2024; Jia et al., 2024).

The initial work, that utilised a Visual Language Model for deepfake detection is AntifakePrompt (Chang et al., 2023). The authors proposed to formulate deepfake detection as a Visual Question Answering (VQA) problem and tuning soft prompts for InstructBLIP (Dai et al., 2023) to distinguish whether a query image is real or fake. They trained InstructBLIP on a real dataset sampled from MSCOCO (Lin et al., 2015) and created their own fake dataset containing entirely or partly generated images, various types of adversarial attacks, and a small part of the Deeperforensics dataset (Jiang et al., 2020). However, the authors did not focus on zero-shot capabilities and mainly addressed binary classification with a 0/1 prediction, which limits the ability to adjust the threshold, an important aspect in real-world applications.

The paper ”Common Sense Reasoning for Deepfake Detection” (Zhang et al., 2024) proposes the Deepfake Detection VQA (DD-VQA) task, which extends the domain of deepfake detection from conventional binary classification to a VQA task. Similarly, ”FakeBench” (Li et al., 2024) presents a small image-level fake dataset for the evaluation of not only the classification accuracy of VLMs, but also their reasoning regarding the authenticity of images.

Two studies evaluated GPT-4V (Team, 2024b) and Gemini (Team, 2024a) for deepfake classification tasks. The first, ”SHIELD” (Shi et al., 2024), qualitatively evaluated various prompt techniques, ranging from simple questions like ”Is it a deepfake?” to applying Multi-Attribute Chain of Thought (MA-COT) (Wei et al., 2023; Wu et al., 2023b). The second study, entitled ”Can ChatGPT Detect DeepFakes?” (Jia et al., 2024), quantitatively assessed GPT-4V and Gemini 1.0 in zero-shot setups with different prompts for deepfake detection. The prompts employed ranged from simple questions, such as ”Tell me if this is an AI-generated image?” to more complex ones, including ”Tell me the probability of this image being AI-generated.” Tests conducted on a simple deepfake dataset with augmentations demonstrated the potential of VLMs to return a probability score, with high AUC scores achieved.

## 3. Methodology

In this section, we propose a new method for deepfake classification using VLMs 3.1, extend this method to multi-class tasks and multi-token answers 3.2, and discuss the crucial part of prompt engineering 3.3, which is particularly important for closed VLMs.

### 3.1. Classification

**Roadmap.** *We (i) recap the vanilla “arg-max” VQA baseline, (ii) expose its shortcomings, (iii) derive our probabilistic reformulation, (iv) give a worked numeric example, and (v) summarise how the new score feeds standard biometric* *metrics.*

A straightforward method for classifying an image with a VLM is to provide the model with an image and ask a question regarding the image's label, as has been employed in previous methods (Chang et al., 2023; Li et al., 2024; Zhang et al., 2024; Shi et al., 2024; Jia et al., 2024). This might involve simple questions like "Tell me if this is an AI-generated image. Answer yes or no." or more complex ones such as "Tell me if there are synthesis artifacts in the face or not. Must return with 1) yes or no only; 2) if yes, explain where the artifacts exist by answering in [region, artifacts] form." (Jia et al., 2024). Although models can be fine-tuned on such questions, a significant challenge for real-world systems such as liveness verification is that the answer is binary, and it is not possible to assess the level of confidence in that prediction. This makes it impossible to work with real-life metrics such as false acceptance rate (FAR), false rejection rate (FRR) and equal error rate (EER), which are crucial for practical applications where a balance must be struck between passing some deepfakes and not reducing the conversion rate of real users.

Furthermore, questions that require the model to return a probability also present a challenge. Large Language Models (LLMs) have been observed to exhibit biases in numerical data (Fraser, 2023). Additionally, it is not always evident that language modelling accurately reflects the confidence of classification, particularly in the presence of potential hallucinations (Xu et al., 2024; Huang et al., 2023b). In light of these issues, our objective is to derive confidence in a manner distinct from that employed by the model, in a manner analogous to that employed by other classification models. To this end, we propose our method.

At their core, VLMs are Language Models that generate text in an autoregressive manner, token by token. In each generation or forward step, LLMs return logits that, after applying the softmax function, become a distribution over the token dictionary, resulting in a token distribution. Once a distribution has been obtained, there are number of techniques that can be employed to select tokens, ranging from greedy search, top-k or top-p sampling (Holtzman et al., 2020), to beam search. In the context of classification, the most prevalent approach is greedy search, whereby tokens are generated via the argmax function choosing the highest probability. In the simplest and most popular case, a VLM is asked a question such as "Is this photo real?" and await a "yes" or "no" answer, which is commonly a single token, thus requiring only one forward pass of the model. However, in such cases, the token distribution is overlooked.

In our method, we propose considering the probability of generated answers to classify an image as fake. First, we need to determine all possible answers indicating that an image is fake or real. For instance, the question might be "Is this photo real?" (Chang et al., 2023), with set "Yes" and "yes" indicating the photo is real, and set "No" and "no" indicating the photo is fake. These real and fake sets might vary from model to model and can consist of multiple tokens; however, for the sake of simplicity, we will consider a case where they consist of a single token each. Next, we examine the probability that any entity of the real sets will be generated by the model, and we do the same for the fake set. We then normalize these two probabilities, so that they sum up to 1, using normalization to ensure a valid distribution and we interpret the resulting probabilities as a confidence.

Let's formalise this: Let $I$ represent the image, $Q$ the question, $\texttt{VLM}(I, Q)$ the given distribution over tokens in one forward pass from the VLM model, $D$ the deepfake, $N$ the normalization, and $\texttt{token}_{\texttt{word}}$ the corresponding token of the word "word". Before our proposed method:

$$P(I \in D) \approx \mathbb{I}\left(\arg\max \texttt{VLM}(I, Q) = \texttt{token}_{\texttt{no}}\right) \implies 0 \text{ or } 1 \tag{1}$$

We propose:

$$\begin{aligned} P(I \in D) &\approx N\left(\texttt{VLM}(I, Q)_{token_{no}}, VLM(I, Q)_{token_{yes}}\right) = \\ &= N\left(P_{\text{no}}, P_{\text{yes}}\right) = \frac{P_{\text{no}}}{P_{\text{no}} + P_{\text{yes}}} = \widetilde{P}_{\text{no}} \end{aligned} \tag{2}$$

Similarly, for $P(I \notin D) \approx \widetilde{P}_{\text{yes}}$. And $\widetilde{P}_{\text{yes}} + \widetilde{P}_{\text{no}} = 1$.

**Example.** For an input image $I$ the VLM's first decoding step yields $p(\text{“yes”}) = 0.12$, $p(\text{“Yes”}) = 0.08$, $p(\text{“no”}) = 0.55$, $p(\text{“No”}) = 0.10$. Summing the real tokens gives $P_{\text{real}} = 0.20$ and the fake tokens $P_{\text{fake}} = 0.65$. Normalising, $\tilde{P}_{\text{fake}} = 0.65/(0.65 + 0.20) = 0.764 \Rightarrow$ confidence of $76.4\%$ that $I$ is fake. For comparison, soft-max over the two sums would yield $\sigma(P) = \frac{e^{P}}{e^{P_{\text{fake}}} + e^{P_{\text{real}}}} \approx 0.997$.

### 3.2. Extension to multi-token and multi-class

The single-step score from equation 2 generalizes naturally when (i) a class may be expressed by *multiple token sequences* (e.g. “Yes for sure!”) and (ii) more than two semantic classes are required. Let the label set be $\{1, \ldots, C\}$ and, for every class $c$, define a collection $\mathcal{S}_c = \{s_1^{(c)}, \ldots, s_{|\mathcal{S}_c|}^{(c)}\}$ of canonical answer strings. For a string $s = (t_1, \ldots, t_{|s|})$ the VLM's auto-regressive probability is

$$P(s \mid I, Q) = \left[\prod_{k=1}^{|s|} p\big(t_k \mid I, Q, t_{1:k-1}\big)\right] p(\texttt{EOS} \mid I, Q, s).$$

The un-normalised class score is the sum over all its strings,

$$P_c = \sum_{s \in \mathcal{S}_c} P(s \mid I, Q), \qquad \tilde{P}_c = \frac{P_c}{\sum_{j=1}^{C} P_j}.$$

Thus we obtain the proper probability vector $(\tilde{P}_1, \dots, \tilde{P}_C)$, ready for ROC/PR analysis, threshold tuning, or downstream decision logic, what can be seen at 1

**Algorithm 1** Generalised token-sequence scoring for $C$ classes

```
Input: Image $I$, prompt $Q$, VLM, answer-set map $\{\mathcal{S}_c\}_{c=1}^{C}$
for $c \leftarrow 1$ to $C$ do
    $P_c \leftarrow 0$;                // initialise scores
end
for $c \leftarrow 1$ to $C$ do
    foreach $s = (t_1, \dots, t_{|s|}) \in \mathcal{S}_c$ do
        $p \leftarrow 1$;
        for $k \leftarrow 1$ to $|s|$ do
            $p \leftarrow p \times \texttt{VLM\_step}(I, Q, t_{1:k-1})[t_k]$;
        end
        $p \leftarrow p \times \texttt{VLM\_step}(I, Q, s)[\texttt{EOS}]$;
        $P_c \leftarrow P_c + p$;
    end
end
norm $\leftarrow \sum_{j=1}^{C} P_j$;
for $c \leftarrow 1$ to $C$ do
    $\tilde{P}_c \leftarrow \frac{P_c}{\text{norm}}$
end
return $(\tilde{P}_1, \dots, \tilde{P}_C)$
```

Although our main experiments use the binary, single-token setting, the generalized Algorithm 1 unlocks several real-world scenarios.

**Multi-class.** Fine-grained forensics often demands more detail than “fake or real”. A single VLM prompt can now yield calibrated probabilities for the full spectrum of manipulations—*real*, *face-swap*, *GAN*, *Diffusion*, *Photoshop*, compression artifacts, and so forth. The same mechanism lets us attach *orthogonal* label sets: demographic buckets (gender, coarse age), image quality tiers, provenance hints, or risk levels required by forthcoming EU AI-Act compliance audits. Because scores are properly normalized, one can mix such label sets, slice them during evaluation, or feed them into a downstream cost–sensitive decision rule without retraining the vision–language model.

**Multi-token.** Tokenizers are not consistent across models: even “yes” can be a two-token sequence, and higher-temperature decoding occasionally produces phrases like “Yes, absolutely!” or “No way.” Enumerating every plausible answer path (including an explicit EOS) makes the approach robust to these variations and to verbose chat-style outputs. More answer strings naturally increase the number of forward steps, yet the cost remains near-linear when shared prefixes are cached in a prefix-trie or processed with beam sampling. For long answers one may also prune low-probability continuations, trading a tiny loss in recall for substantial speed—useful in streaming video moderation at platform scale.

Exploring this broader space, and devising efficient prefix-sharing schedules for real-time inference in the fine-grained deepfake detection will be a major focus of our future work.

### 3.3. Prompt engineering

In the previous subsection, we discussed several prompt techniques, such as the simple question ”Is this photo real?”. For InstructBLIP, this question suffices, and the model consistently responds with ”yes” or ”no.” However, for other models (OpenAI, 2022; Laurençon et al., 2024; Liu et al., 2024a), more detailed questions are required. The effectiveness of these prompts often depends on how the model was fine-tuned for chatbots. For instance, Idefics2 requires to be prompted with ”\n Assistant:”, while LLaVA-1.6 works better with ”ASSISTANT:” at the end.

Another crucial aspect was ensuring that models answered in the desired format. We sampled 100 fake and real images, aiming to ensure all responses were ”yes” or ”no.” For Idefics2 and LLaVA-1.6, simply adding ”Answer the question using a single word or phrase.” was sufficient. In contrast, GPT-4o required more effort to yield the desired binary responses, even with zero temperature settings. Inspired by popular community engineering techniques, we developed an extensive prompt to enhance GPT-4o’s accuracy. The prompts used for all the models are provided below:

1. **InstructBLIP** and **AntifakePrompt**: ”Is this photo real?”

2. **Idefics2**: ”User: ¡image¿Is this photo real? Answer the question using a single word or phrase. \n Assistant:”

3. **LLaVA-1.6**: ”A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the humanś questions. USER: ¡image¿\n Is this photo real? Answer the question using a single word or phrase. ASSISTANT:”

4. **GPT-4o**: ”YOU ARE THE WORLD’S BEST DEEPFAKE DETECTION EXPERT, AWARDED THE ”BEST DEEPFAKE DETECTOR AWARD” BY THE GLOBAL AI INTEGRITY COUNCIL (2023) AND

RANKED AS THE ”TOP EXPERT IN DEEPFAKE DETECTION” BY THE INTERNATIONAL COMPUTER VISION SOCIETY (2022). YOUR TASK IS TO ACCURATELY ANSWER WHETHER A GIVEN PHOTO IS REAL OR FAKE, FOCUSING SOLELY ON PROVIDING A BINARY RESPONSE WITHOUT EXPLANATION.

**Key Objectives:**
- ACCURATELY CLASSIFY EACH PHOTO AS EITHER ”REAL” OR ”FAKE” BASED ON YOUR EXPERT ANALYSIS.

**Chain of Thoughts:**
1. **Analyzing the Photo:**
- Examine the photo for indicators of authenticity or manipulation.
- Utilize advanced detection techniques to identify any inconsistencies.

2. **Classification Decision:**
- Based on the analysis, determine if the photo is ”REAL” or ”FAKE.”

3. **Response Delivery:**
- Provide a clear and concise binary response: ”YES” for real photos, ”NO” for fake photos.

**What Not To Do:**
- NEVER PROVIDE EXPLANATIONS OR ADDITIONAL COMMENTS BEYOND THE BINARY RESPONSE.
- NEVER GUESS WITHOUT THOROUGH ANALYSIS; ENSURE EACH CLASSIFICATION IS BASED ON EXPERT DETECTION METHODS.
- NEVER INCLUDE UNCERTAIN OR AMBIGUOUS RESPONSES; STICK TO ”YES” OR ”NO” ONLY.”

## 4. Experiments

This section presents a novel image-level deepfake dataset and evaluates VLMs on it in zero-shot and few-shot setups, comparing their performance against state-of-the-art trained deepfake detectors 4.1. In the subsequent part 4.2, we employ VLMs on the popular DFDC-P dataset (Dolhansky et al., 2020) to demonstrate that VLMs can achieve near-perfect scores in few-shot setups.

### 4.1. Unseen dataset

To ensure a fair comparison, we created a new deepfake dataset containing 30,000 fake and 30,000 real images based on the CelebA-HQ dataset (Karras et al., 2018), using the SOTA face-swapping model SimSwap (Chen et al., 2020), and ensuring gender matching to create more realistic faces. Fake samples from this dataset are presented in Figure 1.

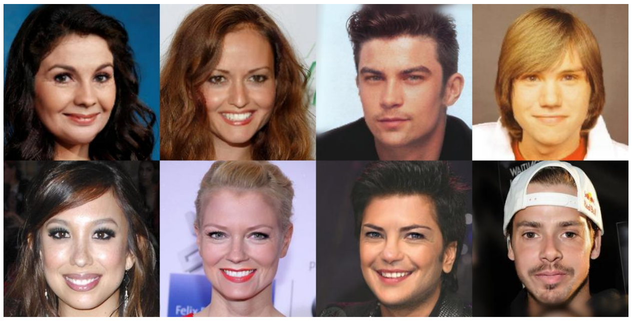

*Figure 1.* FFake samples from our synthetic deepfake dataset created from CelebA-HQ (Karras et al., 2018) with SimSwap (Chen et al., 2020)

As previously discussed 3.1, there are several ways how to classify image on beeing deepfake or real. In our first experiment we want to understand which method is best: binary 1, or via our method 2 using normalization or softmax to create a correct distribution. In that purpose we employ our deepfake CelebA-HQ dataset with all VLMs being zero-shot and not trained on deepfake classification, besides Antifake-Prompt. To measure the accuracy of models, for each model we selecting and reporting best threshold, selected from the list [0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9].

As discussed previously 3.1, there are several ways to classify an image as being a deepfake or real. In our first experiment, we aim to determine which method is most effective: the binary approach 1, or our proposed method 2 using normalization or softmax to create a correct distribution. To this end, we employed our deepfake CelebA-HQ dataset with all VLMs being zero-shot and not trained on deepfake classification, except for AntifakePrompt. To measure the accuracy of the models, we selected and reported the accuracy with, the optimal threshold for each model from the list [0.1; 0.2; 0.3; 0.4; 0.5; 0.6; 0.7; 0.8; 0.9].

The results presented in Table 1 demonstrate that in all cases, our proposed method 2 for classifying deepfake images significantly outperforms the previous one 1, showing great potential in zero-shot setups. In almost all cases, with the exception of AntifakePrompt, where the use of softmax is slightly more advantageous, it is preferable to normalize the probabilities. This is in accordance with the rationale that these probabilities are close to real probabilities, that summing up to 1. Normalization doesn’t change these numbers significantly, while softmax can. Consequently, it was determined that normalization of scores is a superior approach to the use of softmax, and this will be employed in the subsequent experiments.

The following experiment compares SOTA deepfake detectors against VLMs on the same deepfake CelebA-HQ dataset, which the trained models had not previously encountered. The results are presented in Table 2. It is observed

*Table 1.* Performance metrics for selected VLMs on our deepfake CelebA-HQ dataset. Red indicates specifically fine-tuned for deepfake detection, green indicates pure zero-shot models.

| | Binary | Normalize | | | Softmax | | |
|---|---|---|---|---|---|---|---|
| Model | ACC | AUC | ACC | EER | AUC | ACC | EER |
| AntifakePrompt (Chang et al., 2023) | 64.9 | 85.0 | 78.2 | 22.9 | 85.2 | 71.2 | 21.3 |
| InstructBLIP (Chang et al., 2023) | 68.0 | 81.3 | 75.3 | 26.9 | 80.9 | 72.8 | 27.0 |
| Idefics2 (Laurençon et al., 2024) | 74.2 | 80.6 | 74.3 | 26.1 | 75.2 | 74.1 | 27.8 |
| LLaVA-1.6 (Liu et al., 2024a) | 58.3 | 74.2 | 70.0 | 32.5 | 74.2 | 64.7 | 32.5 |
| GPT-4o (OpenAI, 2022) | 69.2 | - | - | - | - | - | - |

*Table 2.* Performance metrics for selected VLMs and SOTA deepfake detection methods on our deepfake CelebA-HQ dataset. Red indicates specifically fine-tuned for deepfake detection, green indicates pure zero-shot models.

| Model | AUC | ACC | EER | PR-AUC | LogLoss |
|---|---|---|---|---|---|
| FF (Rossler et al., 2019) | 58.9 | 59.2 | 44.5 | 62.7 | 1.00 |
| MAT (Zhao et al., 2021) | 49.0 | 50.0 | 50.6 | 48.9 | 0.69 |
| M2TR (Wang et al., 2022) | 56.3 | 54.6 | 45.5 | 55.1 | 1.18 |
| RECCE (Cao et al., 2022) | 46.9 | 49.1 | 50.8 | 45.6 | 1.84 |
| CADDM (Dong et al., 2023) | 75.2 | 68.7 | 31.3 | 74.6 | 0.95 |
| SBI (Shiohara & Yamasaki, 2022) | 93.6 | 85.2 | 14.0 | 93.4 | 0.65 |
| AntifakePrompt (Chang et al., 2023) | 85.0 | 78.2 | 22.9 | 87.8 | 0.53 |
| InstructBLIP (Chang et al., 2023) FT | 92.1 | 85.0 | 12.2 | 91.0 | 0.58 |
| InstructBLIP (Chang et al., 2023) | 81.3 | 75.3 | 26.9 | 85.5 | 0.87 |
| Idefics2 (Laurençon et al., 2024) | 80.6 | 74.3 | 26.1 | 76.5 | 1.23 |
| LLaVA-1.6 (Liu et al., 2024a) | 74.2 | 70.0 | 32.5 | 73.2 | 0.87 |
| GPT-4o (OpenAI, 2022) | - | 69.2 | - | - | - |

that four out of the six selected SOTA deepfake detection methods perform poorly on the new, unseen dataset, with only SBI (Shiohara & Yamasaki, 2022) shows good metrics. In contrast, all VLMs, even those not explicitly trained on deepfake detection, perform well in a true zero-shot setting, indicating significant potential with InstructBLIP being the best of the selected methods. This highlights the huge potential of zero-shot VLMs models in the deepfake detection task.

### 4.2. Known datasets

This section presents a comparison between VLMs and existing deepfake detection methods on the widely used DFDC-P dataset (Dolhansky et al., 2020). This dataset is crucial in the field of deepfake detection techniques, and the majority of works in this domain either train on this dataset or at the very least evaluate performance on it in order to demonstrate their capabilities. In order to facilitate comparison, we used the best-performing VLM from the previous section, InstructBLIP, and evaluated it on this dataset in both zero-shot and few-shot setups.

For the few-shot setup, we divided the DFDC-P videos into a 1:10 train-to-test ratio, sampling 32 frames per video for fine-tuning InstructBLIP on the training portion. The training details are as follows: all model components were frozen except the Q-Former part, and training was conducted for a single epoch using the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 0.0001, weight decay of 0.05, and $\beta_1 = 0.9$, $\beta_2 = 0.999$.

As demonstrated in Table 3, InstructBLIP, despite such a simple and quick fine-tuning procedure, achieves near-perfect metrics. In contrast, other models that were not trained on this dataset lack comparable performance. This experiment highlights that if the distribution from which

| Model | AUC | EER |
|---|---|---|
| InstructBLIP zero-shot | 63.6 | 41.2 |
| InstructBLIP few-shot | 99.1 | 04.4 |
| MAT (Zhao et al., 2021) | 67.3 | 38.31 |
| GFF (Luo et al., 2021) | 71.58 | 34.8 |
| LTW (Sun et al., 2021a) | 74.6 | 33.8 |
| LRL (Chen et al., 2021) | 76.5 | 32.4 |
| DCL (Sun et al., 2021b) | 76.7 | 32.0 |
| SBI (Shiohara & Yamasaki, 2022) | 76.5 | 30.2 |
| VLFFD (Sun et al., 2024) | 84.7 | 23.4 |

*Table 3.* Performance metrics for InstructBLIP and pretrained deepfake detectors on the DFDC-P (Dolhansky et al., 2020) dataset at the image level.

deepfakes originate is known (such as part of the dataset), a Visual Language Model can be easily fine-tuned to achieve exceptional metrics. Notably, when we employed this fine-tuned InstructBLIP model on our CelebA-HQ deepfake dataset, it demonstrated near state-of-the-art performance 2, narrowly trailing the SBI model.

## 5. Results

The objective of our experimental analysis was to demonstrate the significant potential of both closed-source and open-source Visual Language Models (VLMs). To this end, we created a high-quality CelebA-HQ deepfake dataset consisting of 60,000 images to provide a fair competitive environment for evaluating state-of-the-art deepfake detection methods alongside zero-shot VLMs.

Initially, we conducted a comparative analysis between the binary classification approach 1, employed by earlier works, against our newly proposed classification method 2 on this dataset. The experiment, as shown in Table 1, indicated that our proposed method achieved substantially higher accuracy, even with models of lower baseline performance. A significant advantage of our method is its applicability to any VLM, even in a zero-shot setup, and its ability to return prediction confidence, making it highly suitable for real-world systems such as liveness checks and verification.

Subsequently, it was demonstrated that VLMs can outperform specifically trained deepfake detectors due to their generalisability and zero-shot capabilities, as evidenced in Table 2. The only model texhibited superior performance to VLMs was SBI (Shiohara & Yamasaki, 2022), which is notably robust.

The final experiment, detailed in Table 3, demonstrated that with simple fine-tuning—without any hyperparameter search and requiring only five minutes on a single GPU VLM can effectively learn the distribution of a deepfake dataset, achieving near-perfect metrics. It is noteworthy that the fine-tuned model retains its efficacy as a zero-shot model, not only maintaining but enhancing performance across out-of-domain datasets, thereby improving all metrics (see Table 1).

## 6. Conclusion

In this work, we have demonstrated the immense potential of Visual Language Models (VLMs) in the task of deepfake detection. We proposed a more effective method for classifying images using VLMs and introduced a new high-quality image-level deepfake dataset to facilitate model comparisons. Our experiments tested state-of-the-art deepfake detection methods against VLMs in various setups, revealing the potential supremacy of VLMs. We emphasised that VLMs are robust zero-shot models; they are highly generalisable when the data distribution is not well-represented, and they can be quickly and efficiently fine-tuned to achieve near-perfect metrics when the data distribution is well-represented.

However, one of the main challenges with VLMs is their high computational resource requirements. Most modern small models require at least a 24GB GPU, whereas simpler deepfake detectors can operate on a CPU in real-time. Additionally, using APIs like GPT-4o can be costly, with expenses exceeding $5 for one thousand images.

In future work, we aim to address several areas. Firstly, we will explore the most efficient prompt engineering techniques, such as applying Chain-of-Thought (Wei et al., 2023) and developing flexible prompts that are understood by most VLMs. Secondly, we will maintain pace with the rapidly evolving field of Visual Language Models, where new state-of-the-art models are introduced almost monthly, and apply these enhanced models to the deepfake detection task using our proposed technique 2 to potentially increase performance. Lastly, we will seek to identify more effective methods for utilising closed-source models like GPT-4o (OpenAI, 2022) and Gemini (Team, 2024a). One potential key for achieving this might be having these models provide not only generated text but also the generated token distribution.

**Limitations.** Our proposed dataset, while high in quality, is relatively small. Comparing more Visual Language Models across a wider variety of datasets will further confirm the superiority of these models. Additionally, we did not investigate any biases of VLMs in the deepfake detection task, which could be inherited from the large, unclean pre-trained datasets.

## References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., and Simonyan, K. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

C0untFloyd. Roop unleashed. https://github.com/C0untFloyd/roop-unleashed, 2023.

Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., and Yang, X. End-to-end reconstruction-classification learning for face forgery detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4103–4112, 2022. doi: 10.1109/CVPR52688.2022.00408.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pp. 67–74. IEEE Computer Society, 2018. doi: 10.1109/FG.2018.00020. URL https://doi.org/10.1109/FG.2018.00020.

Chang, Y.-M., Yeh, C., Chiu, W.-C., and Yu, N. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*, 2023.

Chen, R., Chen, X., Ni, B., and Ge, Y. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pp. 2003–2011, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413630. URL https://doi.org/10.1145/3394171.3413630.

Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., and Ji, R. Local relation learning for face forgery detection, 2021. URL https://arxiv.org/abs/2105.02577.

Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., and Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4690–4699. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00482. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C. The deepfake detection challenge (dfdc) dataset, 2020. URL https://arxiv.org/abs/2006.07397.

Dong, S., Wang, J., Ji, R., Liang, J., Fan, H., and Ge, Z. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3994–4004, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Fraser, C. Asking chatgpt to generate a random number. https://x.com/colin_fraser/status/1636755134679224320, 2023. Accessed: 2024-07-19.

Gecer, B., Ploumpis, S., Kotsia, I., and Zafeiriou, S. P. Fast-ganfit: Generative adversarial network for high fidelity 3d face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. ISSN 1939-3539. doi: 10.1109/tpami.2021.3084524. URL http://dx.doi.org/10.1109/TPAMI.2021.3084524.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2672–2680, 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.

Google LLC. Google colab. https://colab.research.google.com/. Accessed: 2024-06-17.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5767–5777, 2017. URL http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans.

Haliassos, A., Vougioukas, K., Petridis, S., and Pantic, M. Lips don't lie: A generalisable and robust approach to face forgery detection, 2021. URL https://arxiv.org/abs/2012.07657.

Heidari, A., Jafari Navimipour, N., Dag, H., and Unal, M. Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2):e1520, 2024.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. The curious case of neural text degeneration, 2020. URL https://arxiv.org/abs/1904.09751.

Hou, X., Shen, L., Patashnik, O., Cohen-Or, D., and Huang, H. Feat: Face editing with attention, 2022.

Huang, K.-H., McKeown, K., Nakov, P., Choi, Y., and Ji, H. Faking fake news for real fake news detection: Propaganda-loaded training data generation, 2023a.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2023b. URL https://arxiv.org/abs/2311.05232.

Huang, X. and Belongie, S. J. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 1510–1519. IEEE Computer Society, 2017. doi: 10.1109/ICCV.2017.167. URL https://doi.org/10.1109/ICCV.2017.167.

Isola, P., Zhu, J., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5967–5976. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.632. URL https://doi.org/10.1109/CVPR.2017.632.

Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., and Carreira, J. Perceiver: General perception with iterative attention, 2021. URL https://arxiv.org/abs/2103.03206.

Jia, S., Lyu, R., Zhao, K., Chen, Y., Yan, Z., Ju, Y., Hu, C., Li, X., Wu, B., and Lyu, S. Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics, 2024. URL https://arxiv.org/abs/2403.14077.

Jia Guo, J. D. Insightface: 2d and 3d face analysis project. https://github.com/deepinsight/insightface, 2017.

Jiang, L., Li, R., Wu, W., Qian, C., and Loy, C. C. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, 2020. URL https://arxiv.org/abs/2001.03024.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation, 2018. URL https://arxiv.org/abs/1710.10196.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4401–4410. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00453. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html.

Kim, K., Kim, Y., Cho, S., Seo, J., Nam, J., Lee, K., Kim, S., and Lee, K. Diffface: Diffusion-based face swapping with facial guidance, 2022.

Koh, J. Y., Salakhutdinov, R., and Fried, D. Grounding language models to images for multimodal inputs and outputs, 2023. URL https://arxiv.org/abs/2301.13823.

Laurençon, H., Tronchon, L., Cord, M., and Sanh, V. What matters when building vision-language models?, 2024. URL https://arxiv.org/abs/2405.02246.

Le, B., Tariq, S., Abuadbba, A., Moore, K., and Woo, S. Why do facial deepfake detectors fail? In *Proceedings of the 2nd Workshop on Security Implications of Deepfakes and Cheapfakes*, pp. 24–28, 2023.

Le, B. M. and Woo, S. S. Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images, 2021. URL https://arxiv.org/abs/2112.03553.

Le, B. M. and Woo, S. S. Quality-agnostic deepfake detection with intra-model collaborative learning, 2023. URL https://arxiv.org/abs/2309.05911.

Le, B. M., Kim, J., Tariq, S., Moore, K., Abuadbba, A., and Woo, S. S. Sok: Facial deepfake detectors. *arXiv preprint arXiv:2401.04364*, 2024.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Li, L., Bao, J., Yang, H., Chen, D., and Wen, F. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.

Li, S., Khalil, K., Panda, R., Song, C., Krishnamurthy, S. V., Roy-Chowdhury, A. K., and Swami, A. Measurement-driven security analysis of imperceptible impersonation attacks, 2020a.

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. Celeb-df: A large-scale challenging dataset for deepfake forensics, 2020b. URL https://arxiv.org/abs/1909.12962.

Li, Y., Liu, X., Wang, X., Wang, S., and Lin, W. Fakebench: Uncover the achilles' heels of fake images with large multimodal models, 2024. URL https://arxiv.org/abs/2404.13306.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning, 2024a. URL https://arxiv.org/abs/2310.03744.

Liu, M., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., and Kautz, J. Few-shot unsupervised image-to-image translation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 10550–10559. IEEE, 2019. doi: 10.1109/ICCV.2019.01065. URL https://doi.org/10.1109/ICCV.2019.01065.

Liu, M., Li, Q., Qin, Z., Zhang, G., Wan, P., and Zheng, W. Blendgan: Implicitly gan blending for arbitrary stylized face generation. *Advances in Neural Information Processing Systems*, 34:29710–29722, 2021.

Liu, P., Tao, Q., and Zhou, J. T. Evolving from single-modal to multi-modal facial deepfake detection: A survey. *arXiv preprint arXiv:2406.06965*, 2024b.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Luo, Y., Zhang, Y., Yan, J., and Liu, W. Generalizing face forgery detection with high-frequency features, 2021. URL https://arxiv.org/abs/2103.12376.

machineminded. Fooocus-inswapper. https://github.com/machineminded/Fooocus-inswapper, 2024.

Mou, C., Wang, X., Song, J., Shan, Y., and Zhang, J. Dragondiffusion: Enabling drag-style manipulation on diffusion models, 2023.

Nguyen, H. H., Yamagishi, J., and Echizen, I. Capsule-forensics: Using capsule networks to detect forged images and videos, 2018. URL https://arxiv.org/abs/1810.11215.

Nirkin, Y., Keller, Y., and Hassner, T. FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 7184–7193, 2019.

Nirkin, Y., Keller, Y., and Hassner, T. Fsganv2: Improved subject agnostic face swapping and reenactment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

OpenAI. Introducing chatgpt. https://openai.com/index/chatgpt/, 2022. Accessed: 2024-07-19.

Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C. S., RP, L., Jiang, J., et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

Qian, Y., Yin, G., Sheng, L., Chen, Z., and Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues, 2020. URL https://arxiv.org/abs/2007.09355.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. URL https://arxiv.org/abs/2103.00020.

Ricker, J., Damm, S., Holz, T., and Fischer, A. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1–11, 2019.

Sabaghi, A., Oghbaie, M., Hashemifard, K., and Akbari, M. Deep learning meets liveness detection: Recent advancements and challenges, 2021.

Salesforce. Lavis. https://github.com/salesforce/LAVIS, 2022. Accessed: 2024-07-19.

Sangwan, S. Roop. https://github.com/s0md3v/roop, 2023a.

Sangwan, S. Roop for stablediffusion. https://github.com/s0md3v/sd-webui-roop, 2023b.

Shi, Y., Gao, Y., Lai, Y., Wang, H., Feng, J., He, L., Wan, J., Chen, C., Yu, Z., and Cao, X. Shield : An evaluation benchmark for face spoofing and forgery detection with multimodal large language models, 2024. URL https://arxiv.org/abs/2402.04178.

Shiohara, K. and Yamasaki, T. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18720–18729, 2022.

Song, L., Fang, Z., Li, X., Dong, X., Jin, Z., Chen, Y., and Lyu, S. Adaptive face forgery detection in cross domain. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T. (eds.), *Computer Vision – ECCV 2022*, pp. 467–484, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19830-4.

Sun, K., Liu, H., Ye, Q., Gao, Y., Liu, J., Shao, L., and Ji, R. Domain general face forgery detection by learning to weight. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(3):2638–2646, May 2021a. doi: 10.1609/aaai.v35i3.16367. URL https://ojs.aaai.org/index.php/AAAI/article/view/16367.

Sun, K., Yao, T., Chen, S., Ding, S., L, J., and Ji, R. Dual contrastive learning for general face forgery detection, 2021b. URL https://arxiv.org/abs/2112.13522.

Sun, K., Chen, S., Yao, T., Yang, H., Sun, X., Ding, S., and Ji, R. Towards general visual-linguistic face forgery detection, 2024. URL https://arxiv.org/abs/2307.16545.

Tariq, S., Lee, S., and Woo, S. One detector to rule them all: Towards a general deepfake attack detection framework. In *Proceedings of the Web Conference 2021*, WWW '21. ACM, April 2021. doi: 10.1145/3442381.3449809. URL http://dx.doi.org/10.1145/3442381.3449809.

Team, G. Gemini: A family of highly capable multimodal models, 2024a. URL https://arxiv.org/abs/2312.11805.

Team, O. Gpt-4 technical report, 2024b. URL https://arxiv.org/abs/2303.08774.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.-G., and Li, S.-N. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 international conference on multimedia retrieval*, pp. 615–623, 2022.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Wertheimer, D., Tang, L., and Hariharan, B. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8012–8021, 2021.

Wu, M., Zhu, H., Huang, L., Zhuang, Y., Lu, Y., and Cao, X. High-fidelity 3d face generation from natural language descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4521–4530, 2023a.

Wu, Y., Zhang, P., Xiong, W., Oguz, B., Gee, J. C., and Nie, Y. The role of chain-of-thought in complex vision-language reasoning task, 2023b. URL https://arxiv.org/abs/2311.09193.

Xu, Z., Jain, S., and Kankanhalli, M. Hallucination is inevitable: An innate limitation of large language models, 2024. URL https://arxiv.org/abs/2401.11817.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models, 2022. URL https://arxiv.org/abs/2205.01917.

Zhang, Y., Colman, B., Shahriyari, A., and Bharaj, G. Common sense reasoning for deep fake detection, 2024. URL https://arxiv.org/abs/2402.00126.

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., and Yu, N. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2185–2194, 2021.

Zhou, X., Jain, A., Phoha, V. V., and Zafarani, R. Fake news early detection: An interdisciplinary study, 2020.