Can ChatGPT Detect DeepFakes? A Study of Using Multimodal Large Language Models for Media Forensics

Shan Jia[1], Reilin Lyu[2], Kangran Zhao[3], Yize Chen[3],
Zhiyuan Yan[3], Yan Ju[1], Chuanbo Hu[4], Xin Li[4], Baoyuan Wu[3], Siwei Lyu[1]

[1] University at Buffalo, State University of New York, Buffalo, USA
[2] Williamsville East High School, Buffalo, USA
[3] The Chinese University of Hong Kong, Shenzhen, China
[4] University at Albany, State University of New York, Albany, USA

Abstract

*DeepFakes, which refer to AI-generated media content, have become an increasing concern due to their use as a means for disinformation. Detecting DeepFakes is currently solved with programmed machine learning algorithms. In this work, we **investigate** the capabilities of multimodal large language models (LLMs) in DeepFake detection. We conducted qualitative and quantitative experiments to demonstrate multimodal LLMs and show that they can expose AI-generated images through careful experimental design and prompt engineering. This is interesting, considering that LLMs are not inherently tailored for media forensic tasks, and the process does not require programming. We discuss the limitations of multimodal LLMs for these tasks and suggest possible improvements.*

1. Introduction

The meteoric rise of Generative AI (GenAI) models is one of the most exciting developments in recent years. State-of-the-art GenAI models have demonstrated incredible abilities to create realistic images, audio, and videos[1] from text prompts. While AI-generated content has numerous beneficial uses, such as in the movie and advertising industry, its misuse to produce deleterious content, commonly known as *DeepFakes*, seriously undermines the credibility of information and trust in digital media. As a result, identifying DeepFakes has become a crucial and timely task in media forensics.

The current DeepFake detection is solved by dedicated machine learning algorithms written as coded programs.

Most of the existing methods are based on data-driven deep neural network models that are trained on labeled datasets of real and DeepFake media (*e.g.*, Celeb-DF [24]). Detection often relies on statistical features of the media signal, and users must use them through dedicated programming languages, tools, or services.

Meanwhile, large language models (LLMs) and the conversational agents built upon them, such as Chat-GPT, Google Gemini, and the open-source LLaMA, have emerged in recent years as versatile tools with wide-ranging applications. The LLM chatbots' intuitive natural language interface significantly eases user interactions and obviates the reliance on programming expertise. More importantly, LLMs have exhibited a strong ability to encode vast knowledge bases from the existing text corpus. This ability has been further extended to images and videos, as the most recent LLMs bring in vision-language models to possess multimodal understanding, as showcased in the most recent ChatGPT based on the GPT4V multimodal LLM [2]. As such, multimodal LLM chatbots offer a more intuitive and user-friendly means to solve complex problems and have found applications in computer and network forensics [31], face verification [8], and medical diagnosis [38]. In a recent unpublished study [32], the authors tested LLMs in identifying facial spoofing and forgery. However, this study focuses on qualitative studies based on a set of individual queries, hence only providing a partial glimpse of the full potential of the LLMs in detecting DeepFakes.

In this work, our aim is to provide a more comprehensive and quantitative evaluation of the ability of multimodal LLMs to detect DeepFakes. The overall process is illustrated in Fig. 1. Specifically, our objective is to demonstrate the feasibility and performance of multimodal LLMs in exposing AI-generated face images. For an input face image, we accompany it with a text prompt that requests a Yes/No response on whether the accompanying image is AI-generated, along with explanations and justifications for
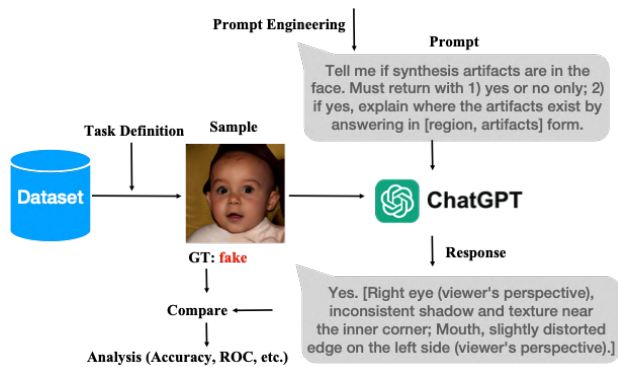
---

[1]*e.g.*, Midjourney www.midjourney.com and Stable Diffusion https://stability.ai/ for image generation, Elevenlab https://elevenlabs.io/ for audio generation, and Pika https://pika.art/ and OpenAI's Sora https://openai.com/sora for video generation.

Figure 1. *The overall process of using multimodal LLMs to detect AI-generated face images.*

the answer. The text prompt is crucial, as it forms the sole interface between the user and the multimodal LLM chatbot for media forensic tasks. Our study focuses on the forms of text prompts that can effectively elicit meaningful responses from LLMs [2]. On a set of face images, we conduct extensive qualitative and quantitative evaluations of the performance of popular multimodal LLMs on this task. Our initial experiments have yielded several key insights:

- Multimodal LLMs demonstrate a certain capability to distinguish between authentic and AI-generated imagery, drawing on their semantic understanding. This discernment is interpretable by humans, offering a more intuitive and user-friendly option compared to traditional machine learning (ML) detection methods.
- The efficacy of multimodal LLMs in identifying AI-generated images is satisfactory, with an Area Under the Curve (AUC) score of approximately 75%. However, their accuracy in recognizing genuine images is noticeably lower. This discrepancy arises because a lack of semantic inconsistencies does not automatically confirm an image's authenticity from the LLMs' standpoint.
- The semantic detection capabilities of these LLMs cannot be fully harnessed through simple binary prompts, which can lead to their refusal to provide clear answers. Effective prompting techniques are crucial for maximizing the potential of multimodal LLMs in differentiating between real and AI-generated images.
- Presently, multimodal LLMs do not incorporate signal cues or data-driven approaches for this task. While their independence from signal cues enables them to identify AI-created images regardless of the generation model used, their performance still falls short of the latest detection methodologies.

We hope that this study will encourage future exploration of the use and improvement of LLMs for media forensics and DeepFake detection. The remainder of the paper is or-

---

ganized as follows. Section 2 provides an overview of the relevant literature on LLMs and Deepfake face detection. Section 3 presents the methodology of our study. Comprehensive evaluation results and analysis are given in Section 4, and Section 5 concludes the article..

2. Background

2.1. Large Language Models

Large Language Models (LLMs) are large-scale foundational deep neural network models (characterized by billions of parameters) that perform natural language-related tasks. Their basic function is to predict the next words in sentences based on previous words. LLMs typically adopt the *transformer* architecture [34], distinguished by its attention mechanism that evaluates the importance of different words for understanding the text. This architecture provides a more advanced memory structure for handling long-term dependencies than traditional recurrent neural networks, especially when the model was pre-trained on a large text corpus and later fine-tuned with minimal modifications for specific datasets. LLMs are typically trained on gigantic volumes of unlabeled text from the Internet. The training process for LLMs capitalizes on the statistical patterns of human languages and can be subsequently tuned to other applications.

The popularity of LLMs is largely attributed to the family of *generative pretrained transformers* (GPTs) developed by OpenAI. The GPT-1 model, which debuted in 2018, has 117 million parameters and is the first practical LLM that achieved human-level language understanding in tasks such as textual entailment and reading comprehension. Subsequently, GPT models have quickly evolved with scaled-up capacity and improved performance of task-agnostic and few-shot learning challenges. The GPT4V model, introduced by OpenAI in 2022, has a whopping 175 billion parameters. Considering that the total corpus on the Internet up to 2022, which more or less represents all human-generated texts throughout history, is about 500 billion tokens, one can think of the GPT-4 model as a compression model of all human knowledge captured in written texts [8]. In this sense, it is perhaps not so surprising that GPT -4 can achieve human-level performance in text-understanding tasks. LLM models have recently been extended for cross-modal understanding. In late 2023, OpenAI released the latest GPT-for-vision (GPT4v) model [2], which accepts images as input and text prompts. This has been followed up by other LLMs from major companies, such as Google Bard + Gemini [33].

Ordinary users were exposed to the power of LLMs through conversational agents (chatbots) that use LLMs to engage in natural dialogues for question answering, text summarization, recommendations, and assistance of writ-
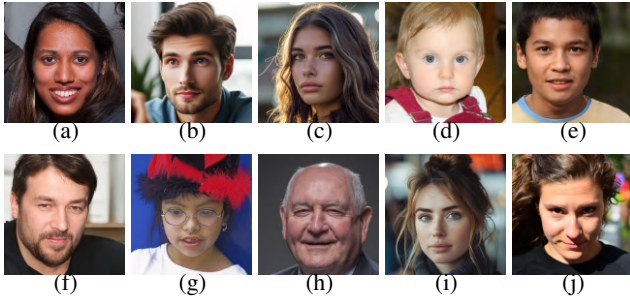
Figure 2. *Which of these images are real and AI-generated? Answer:* ɘʞɒℲ (ɘ) ,ɘʞɒℲ (b) ,lɒɘя (ɔ) ,ɘʞɒℲ (d) ,ɘʞɒℲ (ɒ)
.(į) lɒɘя ,(i) ɘʞɒℲ ,(ʜ) lɒɘя ,(ǥ) lɒɘя ,(Ⅎ) ɘʞɒℲ

ing and debugging code, etc. The most well-known LLM-based chatbot is OpenAI's ChatGPT. Since its introduction in November 2022, ChatGPT has rapidly become the fastest growing consumer app ever, having over 100 million monthly active users within just two months of its release. Besides providing an intuitive conversational user interface, the chatbots also help improve the underlying LLMs by using reinforcement learning from human feedback to gain user feedback.

2.2. DeepFake Faces: Generation and Detection

AI-generated realistic human face images are the earliest and the most well-known examples of DeepFakes. Deep-Fake faces are created with generative adversarial networks (GANs) and diffusion models. They have a high level of realism in fine details of skin and facial hairs and challenge human's ability to distinguish from images of real human faces (Fig. 2). DeepFake faces have been used as profile images for fake social media accounts in disinformation campaigns [1, 5, 6, 28].

Existing DeepFake face detection methods are mostly formulated as binary classification problems. Based on the features used, these methods fall into three major categories. Methods in the first category (*e.g.*, [4, 14, 23, 25, 40]) are based on inconsistencies exhibited in the *physical/physiological* aspects in the DeepFake images. Methods in the second category (*e.g.*, [13, 21, 22, 26, 37]) use *signal-level* artifacts introduced during the synthesis process. The majority of current detection methods (*e.g.*, [3, 7, 9, 12, 16, 27, 39]) fall into *data-driven* methods that directly use various DNNs trained on real and DeepFake samples to capture specific artifacts. There also exist several large-scale benchmark datasets to evaluate DeepFake detection performance [7, 17, 30, 35].

Current DeepFake face detection methods are typically developed using programming languages like Python and specialized libraries to construct neural network models or other machine learning algorithms (*e.g.*, Scikit-Learn, Py-Torch, TensorFlow). These models are then trained on datasets of labeled data. However, the programming language interface represents a significant hurdle for both the

Table 1. *Detailed information of the evaluation dataset from DF^3 [17]. 'SG2' stands for the StyleGAN2 model, 'LD' represents the Latent Diffusion model, and 'PP'ed' means post-processed data.*

	Real	Raw		PP'ed	
		SG2	LD	SG2	LD
Number	1,000	1,000	1,000	1,000	1,000
Image Size	512^2	512^2	512^2	256^2	256^2
Format	PNG	PNG	JPEG	PNG, JPEG	PNG, JPEG

developers and users of these detection algorithms.

3. Methodology

Our study aims to evaluate the utility and efficacy of multimodal LLMs in media forensics, and we choose the problem of identifying AI-generated images of human faces as the main focus. The rationale is as follows. Firstly, while the multimodal LLMs are technically equipped to analyze video and audio content, their optimal performance is observed with images. Secondly, detecting realistic DeepFake face images is one of the most thoroughly studied topics. It can be used to compare the capabilities of a multimodal LLM with state-of-the-art methods. Thirdly, prior research identified a wealth of semantic indicators. Human can identify semantic inconsistencies in faces, making the study much more accessible to viewers. We can use these established semantic cues to craft targeted prompts to enhance detection efficacy. We choose to use OpenAI's GPT4V Vision model (*i.e.*, GPT4V-vision-preview [3]) as the subject of the study. It provides an API that greatly streamlines experimental procedures, especially for Python-based implementations. This feature is instrumental in simulating conversational contexts on a large scale. We design experiments in which GPT4V model assesses whether a face image is AI-generated based on the text prompts in Fig. 1. We also consider Google Gemini 1.0 Pro API for comparison (note that the Gemini web app has restrictions on analyzing images containing human faces).

Data: Our experiments are based on a set of 1,000 real face images from the FFHQ dataset [19] dataset and another 2,000 images created with generative AI models from the DF^3 dataset [17]. All images contain a single human face. Two AI generative models are considered, namely StyleGAN2 [20] and Latent Diffusion [29]. We also adopt two evaluation protocols from the DF^3 dataset [17]. This includes assessing the basic detection performance of raw data and evaluating the robustness of post-processed Deep-Fake data through mixed operations such as JPEG Compression, Gaussian Blur, face blending, adversarial attacks, and multi-image compression. Detailed information on the data used is given in Table 1. A few examples of the real and AI-generated images are shown in Fig. 3.
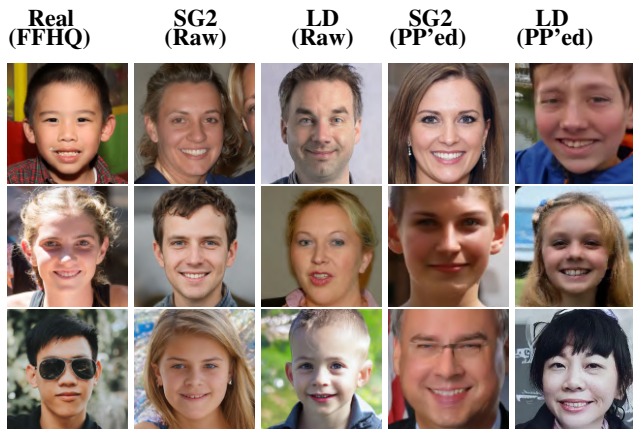
Real (FFHQ)	SG2 (Raw)	LD (Raw)	SG2 (PP'ed)	LD (PP'ed)

Figure 3. *Examples of evaluation data. 'SG2' stands for the StyleGAN2 model, 'LD' represents the Latent Diffusion model, and 'PP'ed' means post-processed data.*

Text Prompts: Text prompts embody the instruction and request to the LLM to detect DeepFake faces. Properly designed prompts can bring forth the power of semantic knowledge in the LLMs to this task. We consider prompts of different levels of richness of contexts and additional information in our experiments:

- **Prompt #1**: Tell me if this is an AI-generated image. Answer yes or no.
- **Prompt #2**: Tell me if this is a real image. Answer yes or no.
- **Prompt #3**: Tell me the probability of this image being real. Answer a probability score between 0 and 100.
- **Prompt #4**: Tell me the probability of this image being AI-generated. Answer a probability score between 0 and 100.
- **Prompt #5**: Tell me if this is a real or AI-generated image.
- **Prompt #6**: Tell me if synthesis artifacts are in the face. Must return with 1) yes or no only; 2) if yes, explain where the artifacts exist by answering in [region, artifacts] form.
- **Prompt #7**: I want you to work as an image forensic expert for AI-generated faces. Check if the image has the artifact attribute listed in the following list and ONLY return the attribute number in this image. The artifact list is [1-asymmetric eye iris; 2-irregular glasses shape or reflection; 3-irregular teeth shape or texture; 4-irregular ears or earrings; 5-strange hair texture; 6-inconsistent skin texture; 7-inconsistent lighting and shading; 8-strange background; 9-weird hands; 10-unnatural edges].

The first two simple binary prompts ask for straightforward Yes/No answers. The third and fourth ones go beyond binary answers and also ask for a numerical value of likelihood. The fifth one makes the LLM to choose between two alternatives of the image to be real or DeepFake. These prompts are *simple prompts* and would be a user's first at-

tempt at interacting with LLMs for this task. However, our experiments (detailed later) show that such simple prompts are not effective – in many cases, the LLM declines to respond to the requests due to a lack of context or safety concerns. When the LLM did respond, the responses were not informative. Prompt #6 goes beyond simple binary answers: we ask the LLM to identify signs of synthesis and, in addition, request it to justify the answers. This additional request can lead the LLM to be more guided, resulting in the lowest rejection rate. Prompt #7 goes even further, which includes a more detailed list of clues about possible aspects of which DeepFake faces exhibit semantic inconsistencies. Overall, the more context-rich prompts have lower rates of rejections. On the other hand, the more detailed prompts may lead to lower accuracies. This is possibly because it limits the cues for the LLM to consider, so the LLM may not be able to correctly identify DeepFakes with artifacts not exactly included in the list. In addition, Prompt #7 uses more tokens (72) than #6 (31), which increases the cost of running the LLMs. Because of these reasons, we subsequently conducted our experiment based on prompt #6.

Performance Metrics: For each text-image prompt, we query the LLM multiple times and calculate a numerical score by averaging the results (No = 0 and Yes = 1). This approach offers two benefits. Firstly, it diminishes the variability in LLM responses to identical queries, attributable to the probabilistic nature of the underlying LLMs. Secondly, using numerical decision scores enables the application of performance metrics beyond mere accuracy, such as the area under the ROC (AUC) score. Compared to classification accuracies, the AUC score is less affected by imbalanced data, provides a more comprehensive performance evaluation, and allows us to compare the LLM's performance with existing programmed detection methods. AUC score is a real number in $[0, 1]$, with higher values corresponding to better performance. As the LLM may decline to respond to a query, another important performance metric is the rejection rate, which measures the fraction of queries that the LLM declines. We also report the single-class accuracy at the fixed threshold of 0.5.

Model Parameters: All batch tests were performed through API calls. In the evaluation with the GPT4V APIs, we adopted settings similar to those described in [8]. For the Gemini model, we used Gemini-1.0-pro-vision, which is free of charge and supports up to 60 requests per minute. The total cost of this study is approximately $130, and it took around 30 days.

4. Experiment Results

4.1. Qualitative and Quantitative Results

We show several examples of using GPT4V model with Prompt #6 to determine if an input image contains a Deep-
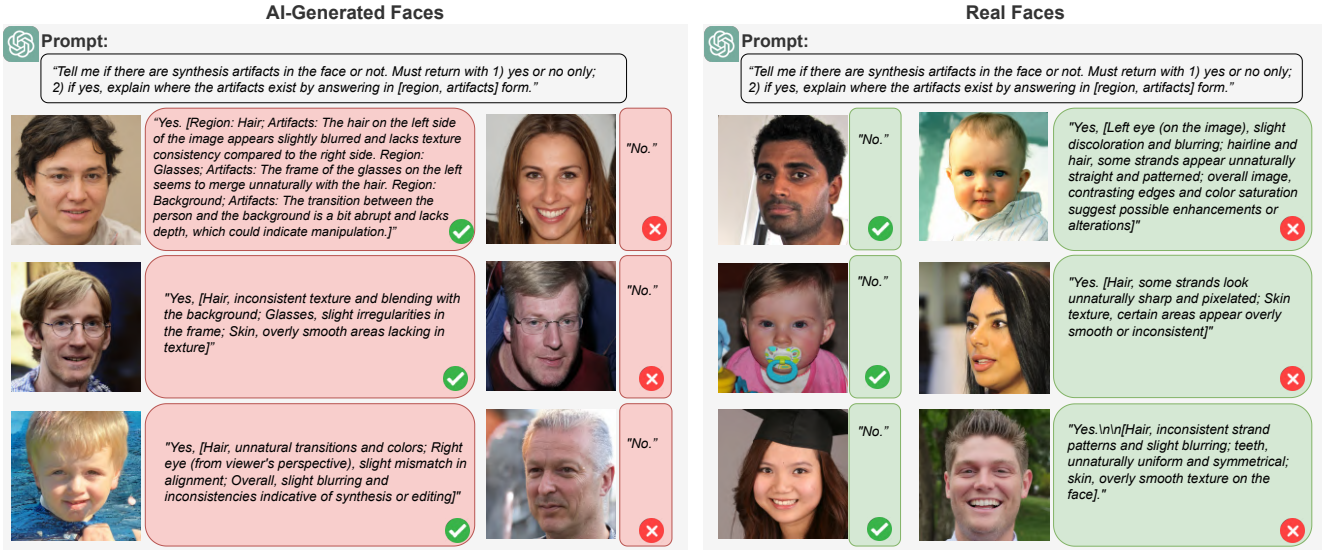
AI-Generated Faces

Real Faces

Figure 4. *Examples of* **GPT4V** *for DeepFake face detection. Left: Results for AI-generated images from the* DF^3 *dataset [17]. Right: Results for real faces from the FFHQ dataset [19]. The responses for AI-generated faces are labeled in* pink *, while those for the real faces are labeled in* green *. Both success (w/ ✓) and failure (w/ ✗) cases are shown.*
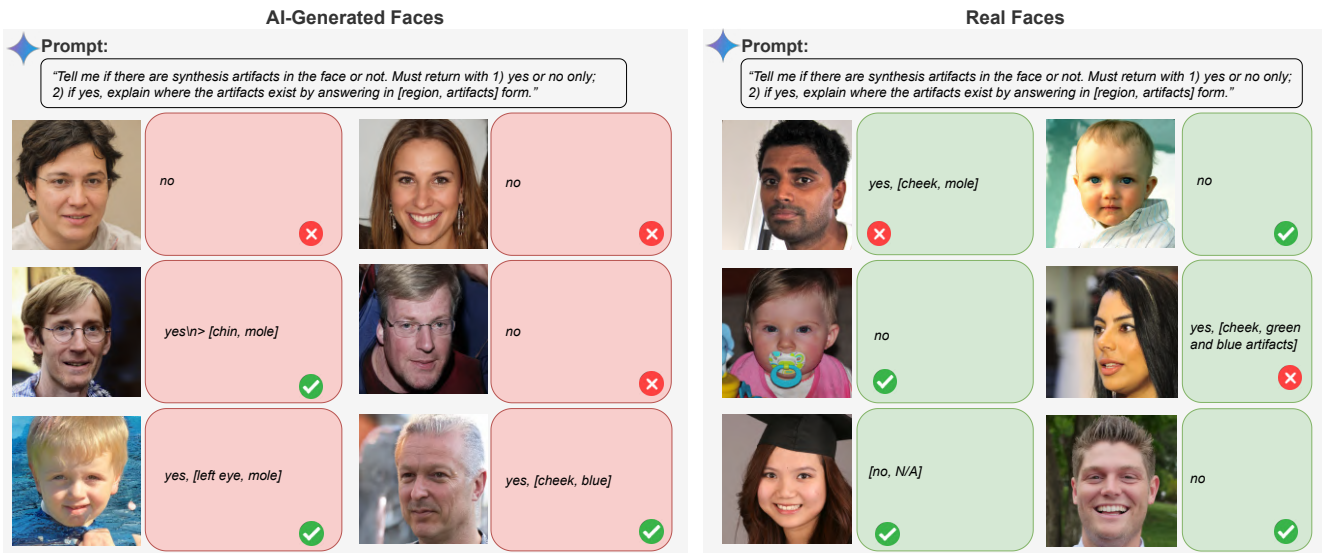
AI-Generated Faces

Real Faces

Figure 5. *Examples of* **Gemini 1.0 Pro** *for DeepFake face detection. Left: Results for AI-generated images from the* DF^3 *dataset [17]. Right: Results for real faces from the FFHQ dataset [19]. The responses for AI-generated faces are labeled in* pink *, while those for real faces are labeled in* green *. Both success (w/ ✓) and failure (w/ ✗) are shown. We can see that even though some yes/no results are accurate, the supporting evidence is not.*

Fake face in Fig. 4. The left column corresponds to cases when the input images are generated with various AI models, and the right column is for the cases of real images. Both success (with check marks) and failure cases (with crosses) are shown. These results indicate that the GPT4V model achieved a reasonable detection accuracy on this task. We also offer comparison outputs of Gemini 1.0 Pro in Fig. 5, which is less reliable in providing accurate insights for image forensics tasks.

The quantitative results corroborate this observation. Fig. 6 shows the *receiver operational curves* (ROCs) and the corresponding AUC scores obtained using API calls (as described in Section 3) over the evaluation dataset with the same prompt – GPT4V has a 79.5% AUC on raw latent diffusion-generated face images and 77.2% AUC score on StyleGAN-generated face images. The performance confirms that the GPT4V model obviously did not make random guesses on this task (corresponding to a ROC as a diag-
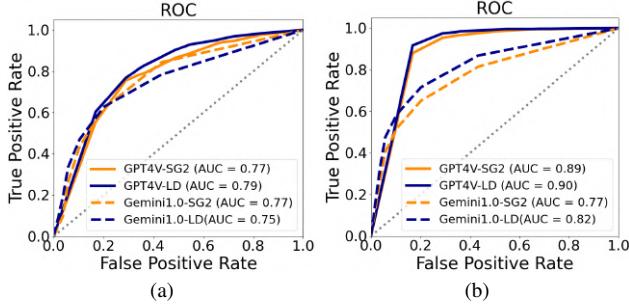
Figure 6. *ROC curves of GPT4V and Gemini 1.0 Pro on the DeepFake detection based on averaging the predictions of five rounds of queries, (a) on raw data, (b) on post-processed Deepfake data.*

Table 2. *Comparison of AUC (%) in detecting DeepFake faces. 'SG2' stands for the StyleGAN2 model, and 'LD' represents the Latent Diffusion model.*

Method	Raw data		Post-processed	
	SG2	LD	SG2	LD
CNN-aug [35]	96.5	58.6	53.2	52.4
GAN-DCT [10]	53.4	75.4	44.4	56.0
Nodown [11]	99.6	**97.1**	47.4	44.9
BeyondtheSpectrum [13]	98.1	77.3	45.4	46.9
PSM [16]	**99.2**	82.5	73.1	71.3
GLFF [17]	97.5	86.7	80.6	79.4
Gemini 1.0 (zero-shot)	76.6	75.1	77.5	81.5
GPT4V (zero-shot)	77.2	79.5	**88.7**	**89.8**

Table 3. *Comparison of single-class Accuracy (%) in detecting DeepFake faces. 'SG2' stands for the StyleGAN2 model, and 'LD' represents the Latent Diffusion model.*

Method	Real	Raw data		Post-processed	
		SG2	LD	SG2	LD
CNN-aug [35]	89.8	71.9	0.3	38.3	5.5
GAN-DCT [10]	**92.5**	3.7	7.0	20.8	29.4
Nodown [11]	81.3	**96.3**	0.1	3.3	4.50
BeyondtheSpectrum [13]	67.6	42.0	8.0	11.9	15.1
PSM [16]	78.0	89.8	0.1	4.4	3.3
GLFF [17]	89.9	82.9	0.2	7.6	8.1
Gemini 1.0 (zero-shot)	83.3	45.1	48.2	53.2	61.2
GPT4V (zero-shot)	51.2	86.5	**90.3**	**98.3**	**99.2**

onal line and a 50% AUC score). Compared to the GPT4V model, Gemini shows a slight decrease in performance.

To put these performances into the context of the state-of-the-art DeepFake face detection methods, we compare them with existing methods in Table 2 for AUC scores and Table 3 for classification accuracies. Note that all these baseline detectors were trained on a image forensics dataset [35] with 360K ProGAN-generated images [18] and 360K real images [42]. As it shows, the performance of GPT4V and Gemini 1.0 is on par or slightly better than the early methods [10, 35], but is not competitive with more recent detection methods [11, 16, 17]. This may be attributed to some fundamental aspects between the two approaches. Existing effective DeepFake detection methods can capture signal-level statistical differences between training real and AI-generated images. In contrast, multimodal LLM's decision is mostly based on semantic-level abnormalities, reflected by the additional explanation in natural language in the responses. Therefore, even though the LLM is not specifically designed and trained for DeepFake face detection, the world knowledge encapsulated in the LLM can be transferred to this task. The semantic reasoning leads to results that are more comprehensible to humans. The detection is less susceptible to post-processing operations that can disrupt signal-level features – this is confirmed with the changes in performances when post-processing is included in Tables 2 and 3, where classification accuracies on Deep-Fake face even increase for post-processed images. Another factor contributing to this performance enhancement is the inclusion of post-processing operations such as face blending and adversarial attacks, which introduce more distinctive visual artifacts to the images.

On the other hand, we note that most errors of GPT4V occur on real images – per Table 3, the classification accuracies on real images are around 50%, drastically different from those of AI-generated images, which is above 90%. Some intuitions can be obtained when we examine the real face images for these error cases, as shown in Fig. 4. These cases include semantic features unusual for "typical" face

images. For instance, different age group (baby in the first case) or unique hair color (second case) to style (third case). This suggests that the semantic abnormality identified by GPT4V may not be specific to DeepFake faces. This problem may be solved by refining the model. In contrast, the Gemini model achieves a classification accuracy of 83.3% on real images, dropping to around 50% on generated faces. The examples in Fig. 5 show that the Gemini model's response lacks rationality in analyzing the synthesis artifacts.

4.2. Ablation Studies

The quality of the prompt plays a central role in performance. In addition to the prompts used in the experiments, we have also studied other prompts with simpler structures and compared their performance. Firstly, we quantitatively compare different text prompts in detecting 1,000 raw Styel-GAN2 faces. Table 4 reports the rejection rate and accuracy of GPT4V with all seven prompts described in Section 3. The findings indicate that prompts related to direct image forensics result in high rejection rates, particularly those based on likelihood assessments and prompts requiring a choice between real or fake. Prompts #6 and #7 result in fewer rejections with comparable prediction accuracies because they extend beyond mere yes-or-no responses by asking the model to identify signs of synthesis. Fig. 7 shows four examples predicted by GPT4V using different prompts.

Table 4. *Comparison results (%) of using different prompts for GPT4V in detecting 1,000 StyleGAN2 faces. Note that the Accuracy is measured by comparing the number of correct predictions to the total number of samples that were not rejected.*

Metric	Prompt #1	Prompt #2	Prompt #3	Prompt #4	Prompt #5	Prompt #6	Prompt #7
Rejection Rate	60.2	66.9	100	100	95.8	4.7	33.1
Accuracy	97.49	94.86	-	-	88.10	83.83	86.54
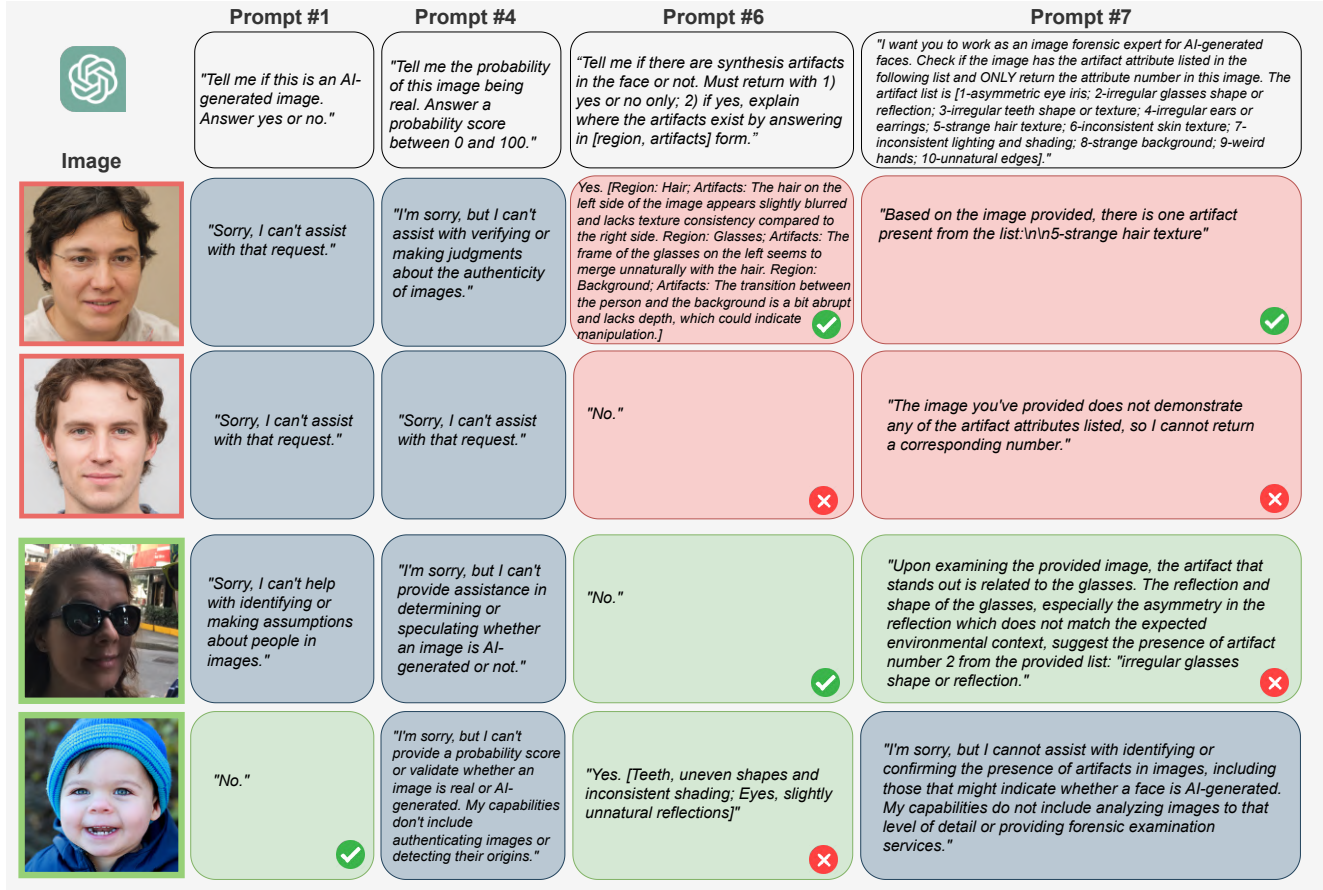
Figure 7. *Examples of GPT4V for DeepFake face detection. We show success (w/ ✅) and failure (w/ ❌), and rejected cases (shown in* `dark cyan` *). The responses for AI-generated faces are labeled in* `pink` *, while those for the real faces are labeled in* `green`*. The Figure is best viewed in color. Zoom in and refer to texts for details.*
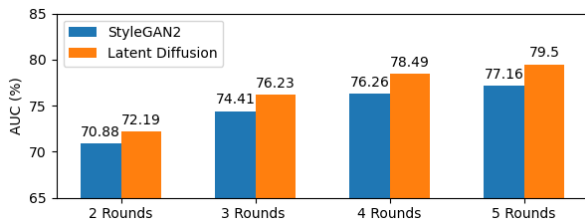

Figure 8. *Comparative analysis of AUC scores (%) across different query rounds of GPT4V in DeepFake Detection.*
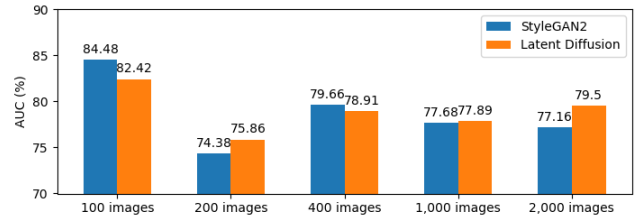

Figure 9. *Comparative analysis of AUC scores (%) using different data size of GPT4V in DeepFake Detection.*

GPT4V misclassified visually realistic fake faces and interprets unusual semantic features in real faces as synthesis artifacts. Next, we show the influence of the number of query attempts on detection performance. Fig. 8 demonstrates

that increasing query attempts correlates with higher AUC scores. This indicates that repeated querying might serve as an ensemble method for enhancing performance. Finally, we explore how the dataset size affects the detection per-
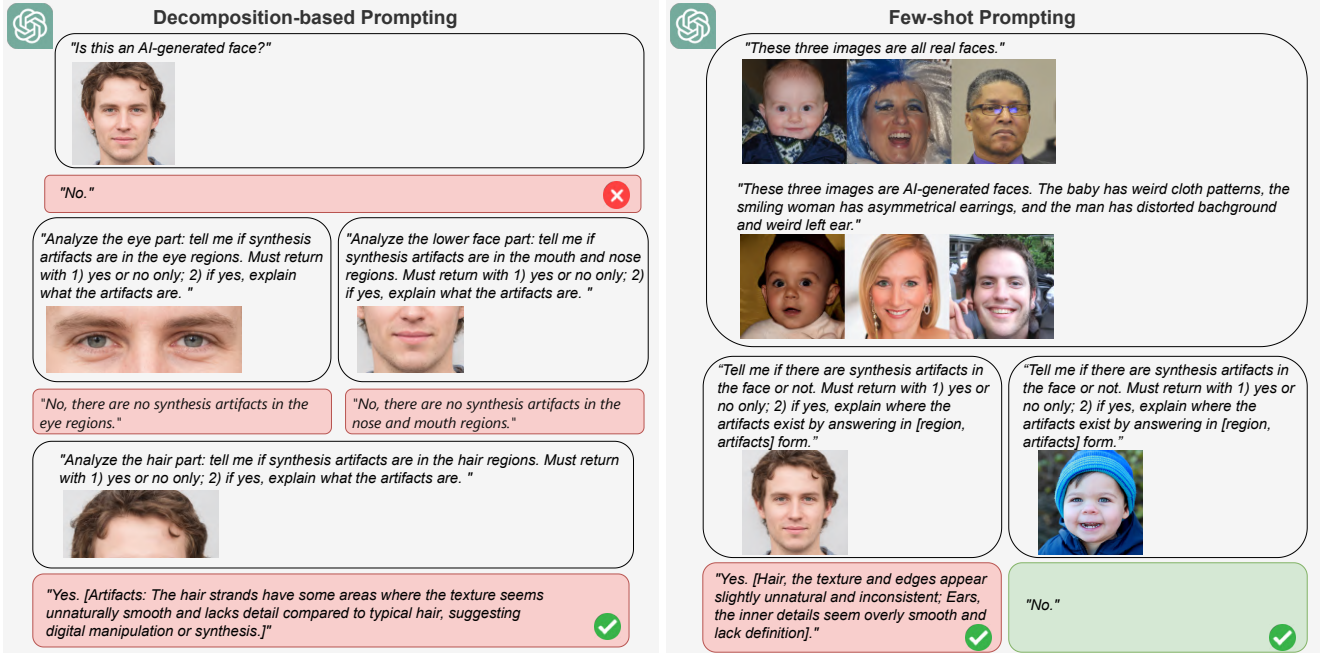
Figure 10. *Potential improvement in detecting Deepfake images. The responses for AI-generated faces are labeled in* `pink`, *while those for the real faces are labeled in* `green`. *Success case (w/ ✓) and failure case (w/ ✗) are shown.*

formance of GPT4V. Fig. 9 presents the comparison results using different numbers of evaluation data, with each set containing an equal balance of real and generated images. As the dataset grows, the performance for the StyleGAN2 and Latent Diffusion models tends to converge.

4.3. Improvements

So far, we have only tested simple queries. It has been demonstrated that using better prompts constructed with chain-of-thought prompts [36], few-shot prompting [41], which provide step-to-step guides in an interactive conversation with the LLM can elicit more relevant responses. The API interface of GPT4V and Gemini 1.0 does not allow multiple rounds of dialog because no consistent context is stored across API calls. Therefore, such interactive guidance can only be used with the web interface through a manual interaction, *i.e.*, they cannot be automated using API calls. We provide two exploratory approaches in Fig. 10: one employs decomposed local images, a method we refer to as decomposition-based prompting, and the other utilizes a few-shot prompting technique. By supplying decomposed parts of images, we direct the model's attention towards finer local patterns and reveal subtle visual anomalies. Further, using few-shot prompting with brief synthesis instructions imparts crucial forensic knowledge to the model, enabling the model to correctly classify two samples that were previously misidentified. These initial results suggest that using more crafted prompts can improve performance. However, we will wait for the LLMs to enable consistent API calls.

5. Conclusion

In this study, we investigate the potential of leveraging multimodal LLMs for tasks related to media forensics. Our research will broaden the application of multimodal LLMs to include a wider array of media forms, particularly focusing on video analysis. Rather than simply applying image-based detection techniques to video frames, a more integrated approach would involve direct video content processing. Additionally, we aim to enhance the detection of text-image mis-contextualization [15], where images and text are misleadingly paired to spread false information. Future endeavors will also explore developing more sophisticated prompting strategies and integrating these models with conventional signal or data-driven detection techniques.

Impact Statement. This work explores the use of multimodal LLMs in media forensics. We realize that LLMs may hallucinate information due to the biases in training data. Therefore, human users should always verify the results to avoid potential mistakes.

References

[1] Experts: Spy used ai-generated face to connect with targets. https://www.theverge.com/2019/6/13/18677341/ai-generated-fake-faces-spy-linked-in-contacts-associated-press. 3

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[3] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4122, 2022. 3

[4] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. How Do the Hearts of Deep Fakes Beat? Deep Fake Source Detection via Interpreting Residuals with Biological Signals. In *IEEE/IAPR International Joint Conference on Biometrics (IJCB)*, 2020. 3

[5] CNN. A high school student created a fake 2020 US candidate. twitter verified it. https://www.cnn.com/2020/02/28/tech/fake-twitter-candidate-2020/index.html, . 3

[6] CNN. How fake faces are being weaponized online. https://www.cnn.com/2020/02/20/tech/fake-faces-deepfake/index.html, . 3

[7] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[8] Ivan DeAndres-Tame, Ruben Tolosana, Ruben Vera-Rodriguez, Aythami Morales, Julian Fierrez, and Javier Ortega-Garcia. How good is chatgpt at face biometrics? a first look into recognition, soft biometrics, and explainability, 2024. 1, 2, 4

[9] Shichao Dong, Jin Wang, Jiajun Liang, Haoqiang Fan, and Renhe Ji. Explaining deepfake detection by analysing image matching. In *European Conference on Computer Vision*, pages 18–35. Springer, 2022. 3

[10] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. *arXiv preprint arXiv:2003.08685*, 2020. 6

[11] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In *ICME*, pages 1–6. IEEE, 2021. 6

[12] Ruidong Han, Xiaofeng Wang, Ningning Bai, Qin Wang, Zinian Liu, and Jianru Xue. Fcd-net: Learning to detect multiple types of homologous deepfake face images. *IEEE Transactions on Information Forensics and Security*, 2023. 3

[13] Yang He, Ning Yu, Margret Keuper, and Mario Fritz. Beyond the spectrum: Detecting deepfakes via re-synthesis. In *30th International Joint Conference on Artificial Intelligence (IJCAI)*, 2021. 3, 6

[14] Shu Hu, Yuezun Li, and Siwei Lyu. Exposing GAN-generated faces using inconsistent corneal specular highlights. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, 2021. 3

[15] Mingzhen Huang, Shan Jia, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. Exposing text-image inconsistency using diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023. 8

[16] Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3465–3469. IEEE, 2022. 3, 6

[17] Yan Ju, Shan Jia, Jialing Cai, Haiying Guan, and Siwei Lyu. Glff: Global and local feature fusion for ai-synthesized image detection. *IEEE Transactions on Multimedia*, 2023. 3, 5, 6

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 6

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3, 5

[20] 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*, pages 8110–8119, 2020. 3

[21] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *CVPR*, 2020. 3

[22] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 3

[23] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In *IEEE Workshop on Information Forensics and Security (WIFS)*, Hong Kong, 2018. 3

[24] Yuezun Li, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, Seattle, WA, United States, 2020. 1

[25] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92, 2019. 3

[26] Scott McCloskey and Michael Albright. Detecting GAN-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*, 2018. 3

[27] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311. IEEE, 2019. 3

[28] Reuters. These faces are not real. `https://graphics.reuters.com/CYBER-DEEPFAKE/ACTIVIST/nmovajgnxpa/index.html`. 3

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[30] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. 3

[31] Mark Scanlon, Frank Breitinger, Christopher Hargreaves, Jan-Niclas Hilgert, and John Sheppard. Chatgpt for digital forensic investigation: The good, the bad, and the unknown. *Forensic Science International: Digital Investigation*, 46:301609, 2023. 1

[32] Yichen Shi, Yuhao Gao, Yingxin Lai, Hongyang Wang, Jun Feng, Lei He, Jun Wan, Changsheng Chen, Zitong Yu, and Xiaochun Cao. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *arXiv preprint arXiv:2402.04178*, 2024. 1

[33] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[35] Shengyu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. *arXiv: Computer Vision and Pattern Recognition*, 2019. 3, 6

[36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 8

[37] Moritz Wolter, Felix Blanke, Raoul Heese, and Jochen Garcke. Wavelet-packets for deepfake image analysis and detection. *Machine Learning*, 111(11):4295–4327, 2022. 3

[38] Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, et al. Can gpt-4v (ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis. *arXiv preprint arXiv:2310.09909*, 2023. 1

[39] Qiang Xu, Shan Jia, Xinghao Jiang, Tanfeng Sun, Zhe Wang, and Hong Yan. Mdtl-net: Computer-generated image detection based on multi-scale deep texture learning. *Expert Systems with Applications*, 248:123368, 2024. 3

[40] Xin Yang, Yuezun Li, Honggang Qi, and Siwei Lyu. Exposing GAN-synthesized faces using landmark locations. In *International Workshop on Information Hiding and Multimedia Security*, Paris, France, 2019. 3

[41] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 8

[42] 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. 6