APSIPA Transactions on Signal and Information Processing, 2025, 14, e11

Original Paper

How Good is ChatGPT at Audiovisual Deepfake Detection: A Comparative Study of ChatGPT, AI Models and Human Perception

Sahibzada Adil Shahzad[1,2], Ammarah Hashmi[1,3], Yan-Tsung Peng[2], Yu Tsao[3] and Hsin-Min Wang[1*]

[1]*SNHCC, Taiwan International Graduate Program, Institute of Information Science, Academia Sinica, Taipei 11529, Taiwan*
[2]*Department of Computer Science, National Chengchi University, Taipei 11529, Taiwan*
[3]*Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu 300044, Taiwan*
[4]*Research Center for Information Technology Innovation, Academia Sinica, Taipei 11529, Taiwan*

ABSTRACT

Multimodal deepfakes involving audiovisual manipulations are a growing threat because they are difficult to detect with the naked eye or using unimodal deep learning-based forgery detection methods. Audiovisual forensic models, while more capable than unimodal models, require large training datasets and are computationally expensive for training and inference. Furthermore, these models lack interpretability and often do not generalize well to unseen manipulations. In this study, we examine the detection capabilities of a large language model (LLM) (i.e., ChatGPT) to identify and account for any possible visual and auditory artifacts

[*] Corresponding author: whm@iis.sinica.edu.tw

Received 24 January 2025; revised 23 April 2025; accepted 26 May 2025
ISSN 2048-7703; DOI 10.1561/116.20250004

and manipulations in audiovisual deepfake content. Extensive experiments are conducted on videos from two benchmark multimodal deepfake datasets to evaluate the detection performance of ChatGPT and compare it with the detection capabilities of state-of-the-art multimodal forensic models and humans. Experimental results demonstrate the importance of domain knowledge and prompt engineering for video forgery detection tasks using LLMs. Unlike approaches based on end-to-end learning, ChatGPT can account for spatial and spatiotemporal artifacts and inconsistencies that may exist within or across modalities. Additionally, we discuss the limitations of ChatGPT for multimedia forensic tasks.

Keywords: LLM, ChatGPT, deepfake, audiovisual deepfake, multi-modality, video forensics, forgery detection

1 Introduction

Synthetic multimedia content has become both innovative and a significant threat in recent years. Deepfake images and videos created using artificial intelligence (AI) and deep learning (DL) techniques have attracted public and academic attention. This synthetic content is generated by generative adversarial networks (GANs) [16, 36] and more sophisticated AI techniques such as diffusion models [22]. While deepfake technology has many innovative applications in education, entertainment, and other fields [32], it is a double-edged sword that can be used for unethical purposes, such as pornography, political defamation, identity theft, fraud, misinformation, and disinformation [41, 14, 46]. Unethical use of this technology can lead to political instability and social violence [46]. On the one hand, deepfake technology continues to evolve to create more convincing and realistic fake multimedia content. Social media, on the other hand, plays a catalytic role in spreading such content. Therefore, timely detection of deepfake content is crucial to avoid any damage and loss to human society [41].

Audiovisual deepfakes that involve multimodal manipulation are a more convincing type of forgery, with attackers attacking audio, video, or both modalities. Unimodal image forensics method [35, 33], video forgery detectors [1, 37, 34, 17] and spoofed audio detectors [49, 45, 39, 51] are generally unable to identify forgeries across multiple modalities, although they may be good at detecting forgeries in the specific modality they focus on. To address this challenge, the research community has developed sophisticated tools and algorithms to detect audiovisual forgeries in videos. These specialized tools require knowledge of multimedia forensics as well as knowledge of deep learning. Furthermore, these tools do not generalize well to other unseen datasets and manipulations.

Large language models (LLMs) are a major advancement in the field of artificial intelligence. They are trained on a large amount of data and can perform well in various natural language processing (NLP) tasks such as text generation, summarization, classification, completion, sentimental analysis, machine translation, and question answering. Their applications even go beyond the aforementioned NLP tasks and can be used as writing assistants, learning tools, productivity tools, coding assistants, software development, healthcare, legal assistance, entertainment, and more. Despite being primarily designed for NLP tasks, OpenAI's ChatGPT can analyze image, audio, and video content. Taking advantage of its support for multimodal input, we studied the potential and limitations of ChatGPT for audiovisual deepfake detection.

The research questions we aimed to address in this study are as follows:

- Can ChatGPT perform multimedia forensic tasks?
- Is ChatGPT capable of detecting forgery based on artifacts in audio and visual modalities?
- What is the role of prompt engineering in using ChatGPT to detect audiovisual deepfakes?
- Which performs better at identifying forgeries in audiovisual deepfakes, ChatGPT, humans, or AI models?
- How interpretable is ChatGPT for forgery detection?
- What are the limitations of ChatGPT in detecting multimodal deepfakes?

The main contributions of our work are threefold:

- We explore for the first time the potential of ChatGPT for audiovisual forgery detection tasks.
- We compare the performance of ChatGPT with human and state-of-the-art AI models on audiovisual forgery detection tasks.
- We highlight the strengths and limitations of ChatGPT on audiovisual forgery detection tasks.

2 Related Work

The societal impact of synthetic media content has prompted research from multiple angles within the forensic science community. In [27], the authors investigated synthetic content from multiple perspectives such as multimedia content production, representation, media audience dynamics, gender, politics, law, and regulation, and concluded that the intersection between media and deepfake content can have multiple impacts on individuals and society. A study in [57] on the impact of unreliable deepfake information on voter behavior in US elections and democracy suggests multi-stakeholder partnerships and technological approaches for identifying and mitigating manipulated content on public platforms. In [8], the balance between innovation and ensuring fair protection under existing laws is explored, particularly as generative AI blurs the line between human and machine-generated works. The US FDA's regulation of AI in medical devices and the European AI Act, which classify AI applications based on potential harm, are initiatives aimed at addressing challenges and aligning AI-generated content and applications with human-centered values. This analysis is essential for developing a legal framework that addresses the ethical and practical implications of the creativity of generative AI in the legal domain. A recent study in [2] investigated the risk of deepfakes in legal proceedings, where altered audiovisual evidence could compromise the integrity of justice. It highlights how deepfake technology can influence the outcome of cases based on subjective human judgment. The findings point to the need for changes to the legal framework to ensure that key judicial principles such as the presumption of innocence and the right to a fair trial are protected. Furthermore, research on human perception of audiovisual deepfake videos [20] shows that it is difficult for people to accurately distinguish deepfake content from real videos, mainly due to the realistic visual and acoustic manipulations involved.

Audiovisual deepfakes can be broadly categorized into three types, as shown in Figure 1. The first type is “Fake Video Real Audio” (FVRA), in which the visual frames are manipulated using techniques such as Faceswap [31], Fsgan [38], and Wav2lip [40], while the audio modality remains unaltered. The second type is “Real Video Fake Audio” (RVFA), where the video frames are authentic, but the audio modality is manipulated using techniques such as SV2TTS [25], a real-time voice cloning tool that can synthesize fabricated audio content. The third type is “Fake Video Fake Audio” (FVFA), where both modalities are manipulated. In this type, face manipulation can be done using methods like Faceswap and Fsgan, while lip synchronization/manipulation can be achieved using Wav2lip. Additionally, cloned or synthesized audio can be integrated with visually manipulated frames to produce more realistic and convincing deepfake videos.

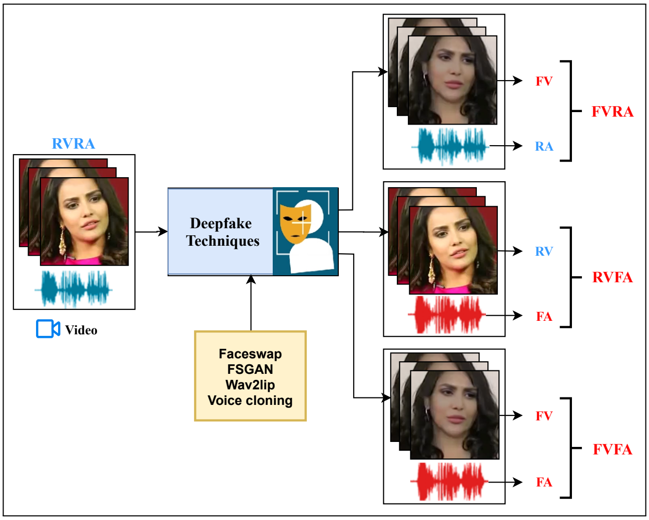


Figure 1: Illustration of audiovisual deepfake manipulations. Original video content is represented as RVRA (real video with real audio. Through deepfake manipulation techniques, three manipulated types are generated: FVRA (fake video with real audio), RVFA (real video with fake audio), and FVFA (fake video with fake audio). Blue text represents the "real modality" of the video content, while red text represents the "fake modality".

The multimedia forensics community has developed several data-driven audiovisual deepfake detection methods based on multimodal feature fusion [59], ensemble learning [18, 28], and synchronization features [42, 43]. Models based on Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Transformers [10, 23, 43, 19, 54, 58, 26] have been widely used to detect forgeries in either modality and are trained on multimodal deepfake datasets. These methods provide a binary output for the input video, indicating whether the input video is genuine or spoofed. Recent work in [60] introduced regularization strategies to better integrate and disentangle audiovisual cues. [9] used facevoice synchronization as discriminative cues to detect audiovisual forgeries. Disadvantages of these end-to-end learning-based methods include reliance on large datasets, heavy training, and lack of interpretability and generalization. Bias, imbalance, and lack of diversity in training data can lead to fairness, generalization, and security issues for detection models [48, 52].

Recently, with the emergence of LLMs, the research community has begun to utilize these models to solve various tasks in different fields, beyond their original purpose. For example, while ChatGPT is primarily designed for NLP tasks, the multimodal mode in ChatGPT-4 enables it to handle multimodal inputs and analyze content from a multimodal perspective [4, 47, 56]. Many studies have investigated the performance of LLMs in various challenging tasks, such as image forensics [24], facial biometrics [11], tampered image detection [55], fake news detection [7], NLP [30], cheap-fake detection [50], global warming [5], education [21], public health [6], and medical applications [53]. These studies highlight the strengths and limitations of LLMs, focusing specifically on ChatGPT's effectiveness in handling these tasks. Unlike traditional machine learning-based multimedia forensic tools, LLMs are readily accessible and can be used for multimedia forgery detection tasks. In this study, we aim to leverage the implicit knowledge embedded in ChatGPT and the generalization ability of multimodal inputs to accomplish the audiovisual deepfake detection task.

3 Methodology

Figure 2 shows an LLM-based approach for multimodal forgery detection, where the text prompts and video with corresponding audio are used as inputs. Based on the given prompts, the model works as a black box and produces multiple analysis results on the input video, such as visual, acoustic, and audiovisual analysis and their corresponding predictions. Our goal is to evaluate the detection capabilities of ChatGPT. This model is trained on multimodal data and can be used for audiovisual forgery detection tasks. Deepfake attacks usually target high-level facial features and voice identities; therefore, we choose frontal face videos with voices to evaluate the detection performance of ChatGPT. Unlike traditional end-to-end models that leverage low-level features, LLMs provide high-level features and descriptions to analyze multimodal inputs. In this study, we used OpenAI's GPT-4 to conduct audiovisual analysis of deepfake videos. Unlike other deep learning-based models, it provides interpretability by explaining the reasoning behind the final decision, thereby increasing the transparency of the decision-making process.

3.1 Prompt Design

Inspired by previous research on LLM-based image forensics [24], we proposed the following custom prompts, ranging from simple, binary-answer prompts to advanced, context-rich prompts designed to account for artifacts and manipulations in audiovisual content:

- P1: Tell me if this is an AI-generated video by analyzing both audio and video modalities. Answer yes or no.

- P2: Tell me if this is a real video by analyzing both audio and video modalities. Answer yes or no.

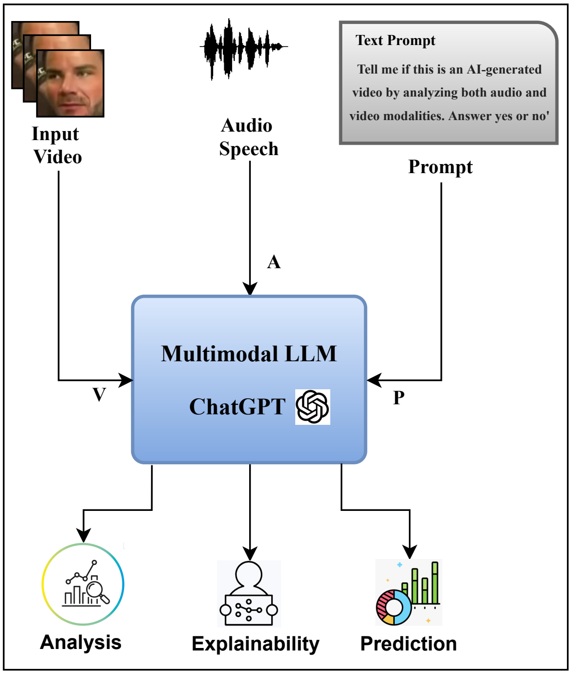

Input Video
Audio Speech
Text Prompt
Tell me if this is an AI-generated video by analyzing both audio and video modalities. Answer yes or no'
Prompt
A
V
P
Multimodal LLM
ChatGPT
Analysis
Explainability
Prediction

Figure 2: Illustration of the multimodal capabilities of a large language model, which takes inputs such as visual frames, audio speech, and text prompts to produce outputs that include audiovisual analysis, interpretation, and authenticity prediction.

- P3: Tell me the probability of this video being real by analyzing both audio and video modalities. Answer a probability score between 0 and 100.
- P4: Tell me the probability of this video being AI-generated by analyzing both audio and video modalities. Answer a probability score between 0 and 100.
- P5: Tell me if this is a real or AI-generated video by analyzing both audio and video modalities.
- P6: Tell me if synthesis artifacts are in the visual frames or audio. Must return with 1) yes or no only; 2) if yes, explain where the artifacts exist by answering in [region, artifacts] form.
- P7: I want you to work as a multimedia forensic expert for audiovisual deepfake detection. Check if the video has the artifact attribute

listed in the following list, and ONLY return the attribute number in this video. The artifact list is [1-unnatural face edges; 2-inconsistent lighting and shading; 3-irregular teeth shape or texture; 4-irregular lip movement; 5-inconsistent skin texture; 6-spectral artifacts; 7-phoneme artifacts; 8-inconsistencies in speech patterns; 9-voice quality issues; 10-lack of synchronization between audio and video].

3.2 Input of LLM Model

We feed the videos directly into the LLM-based ChatGPT model without performing any preprocessing or transcribing the videos for analysis. The model extracts audio from the video and performs visual and acoustic analysis based on input prompts. Let $\mathrm{X} = (\mathrm{x_v}, \mathrm{x_a}, \mathrm{x_t})$ denote the entire input, where $\mathrm{x_v}$ represents the video, $\mathrm{x_a}$ represents the audio, and $\mathrm{x_t}$ represents the custom text prompt. The model outputs its final prediction as:

$$\hat{y} = f_{\mathrm{LLM}}(\mathrm{X}) = f_{\mathrm{LLM}}(\mathrm{x_v}, \mathrm{x_a}, \mathrm{x_t}), \tag{1}$$

where f_{LLM} is the underlying function.

3.3 Audiovisual Analysis

Based on the given input text prompts, the model performs audio analysis by analyzing spectral features, zero crossing rate, mel-frequency cepstral coefficients (MFCC), amplitude envelope, amplitude range, median amplitude, spectral centroid, spectral rolloff, and silence ratio. For video, the model performs analysis such as blurriness, pixelation, lighting, frame difference mean, frame difference standard deviation, unnatural expression, skin texture, lip-syncing, and structural similarity index (SSIM). It also performs multimodal analysis to verify consistency between visual and audio modalities through synchronization checks. By combining unimodal acoustic and visual analyses with multimodal analysis, joint analysis is performed to reach a final prediction or suggest further manual inspection.

3.4 Prediction Assignment

The final prediction $\hat{y}$ for each input video is determined based on the following factors: yes/no response, probability score, overall conclusion, artifact-free versus artifact list, and estimated likelihood of the input video being real or fake. Fake classes are assigned label 1, while label 0 represents real classes.

3.5 ChatGPT vs Human vs AI Models

To understand the detection capabilities of ChatGPT, humans, and AI models, we follow the study in [20], where the authors reported a comparative analysis between humans and deep-learning-based multimodal forensic models. Their results concluded that state-of-the-art AI models surpass humans in detecting multimodal deepfakes. Furthermore, participants often showed overconfidence in their detections, with their average accuracy being lower than their confidence level. To evaluate the detection performance of ChatGPT, we selected the same video subset used in [20] from the FakeAVCeleb [29] and DFDC [12] datasets for a fair comparison.

4 Experiments and Results

4.1 Dataset Selection

Following the study in [20], we selected the same 40 videos from the FakeAVCeleb [29] and DFDC [12] datasets, respectively, as test sets to ensure fair comparison with humans and state-of-the-art multimodal forensic models. Each test set contains 20 real videos and 20 fake videos, and each category contains an equal number of male and female videos.

4.2 Evaluation Metrics

For evaluation, we calculate precision, recall, F1-score, and accuracy, which are defined as,

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{3}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{4}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{5}$$

where TP, TN, FP, and FN stand for True Positive (fake videos correctly detected as fake), True Negative (real videos correctly detected as real), False Positive (real videos incorrectly detected as fake), and False Negative (fake videos incorrectly detected as real), respectively. Additionally, we calculate the rejection rate to evaluate the effectiveness of input text prompts:

$$\text{Rejection Rate} = \frac{\text{Number of Rejected Prompts}}{\text{Total Number of Prompts}} \times 100. \tag{6}$$

4.3 Results

4.3.1 Comparing Different Text Prompts on the FakeAVCeleb Test Set

Table 1 lists the performance of different text prompts (P1-P7 in Section 3.1) evaluated on the FakeAVCeleb test set. Figure 3 shows a bar graph comparison of TP, TN, FP, and FN (see Section 4.2) for different text prompts. Additionally, Figure 4 shows an example of custom text prompts, ChatGPT audiovisual analysis, corresponding predication, and ground truth label. Next, we analyze each text prompt in detail and discuss the results.

Table 1: Precision, Recall, F1 Score, Rejection Rate, and Accuracy for different text prompts (P1-P7) on the FakeAVCeleb test set.

Prompt	Precision	Recall	F1 Score	Rejection Rate	Accuracy
P1	0.583	0.368	0.451	7.50	54.0
P2	0.625	0.250	0.357	2.50	53.8
P3	0.625	0.250	0.357	0.00	55.0
P4	0.571	0.600	0.585	0.00	57.5
P5	0.571	0.200	0.300	0.00	52.5
P6	0.607	0.850	0.706	0.00	65.0
P7	0.600	0.900	0.720	0.00	65.0

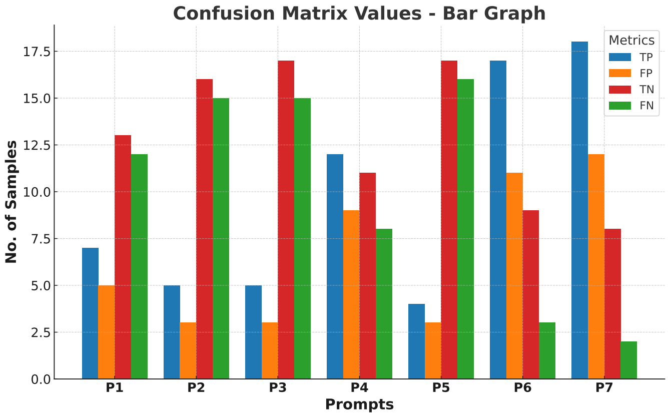
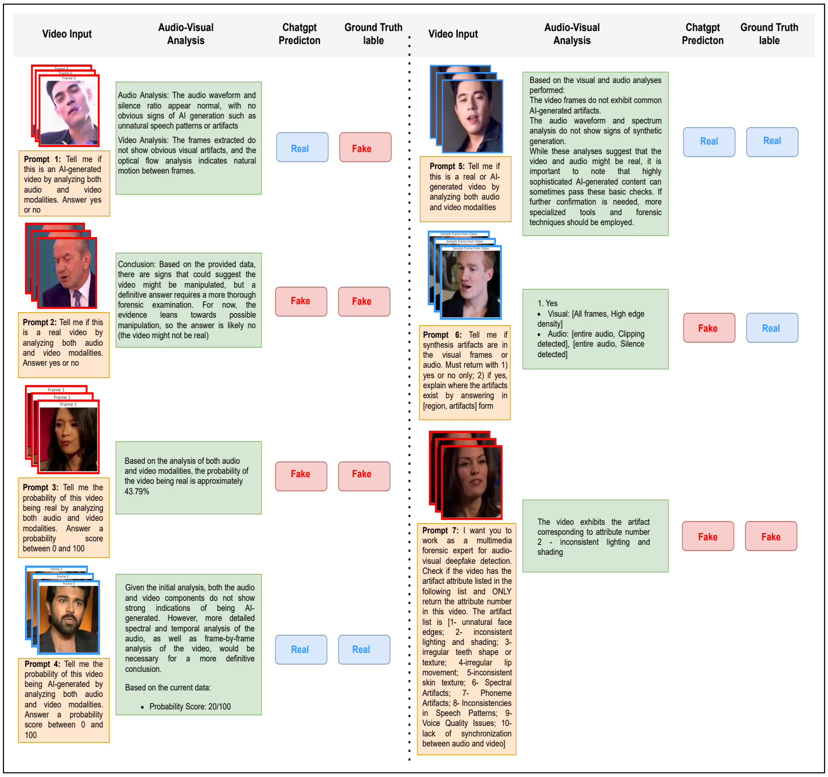


Figure 3: Bar graph comparing the number of True Positives, False Positives, True Negatives, and False Negatives for different text prompts on the FakeAVCeleb test set.

- Prompt P1: As can be seen from Table 1, the accuracy of P1 reaches 54.0%, which is only slightly higher than the 50% of random guessing,



Video Input	Audio-Visual Analysis	Chatgpt Predicton	Ground Truth lable
Prompt 1: Tell me if this is an AI-generated video by analyzing both audio and video modalities. Answer yes or no	Audio Analysis: The audio waveform and silence ratio appear normal, with no obvious signs of AI generation such as unnatural speech patterns or artifacts Video Analysis: The frames extracted do not show obvious visual artifacts, and the optical flow analysis indicates natural motion between frames.	Real	Fake
Prompt 2: Tell me if this is a real video by analyzing both audio and video modalities. Answer yes or no	Conclusion: Based on the provided data, there are signs that could suggest the video might be manipulated, but a definitive answer requires a more thorough forensic examination. For now, the evidence leans towards possible manipulation, so the answer is likely no (the video might not be real)	Fake	Fake
Prompt 3: Tell me the probability of this video being real by analyzing both audio and video modalities. Answer a probability score between 0 and 100	Based on the analysis of both audio and video modalities, the probability of the video being real is approximately 43.79%	Fake	Fake
Prompt 4: Tell me the probability of this video being AI-generated by analyzing both audio and video modalities. Answer a probability score between 0 and 100	Given the initial analysis, both the audio and video components do not show strong indications of being AI-generated. However, more detailed spectral and temporal analysis of the audio, as well as frame-by-frame analysis of the video, would be necessary for a more definitive conclusion. Based on the current data: • Probability Score: 20/100	Real	Real
Prompt 5: Tell me if this is a real or AI-generated video by analyzing both audio and video modalities	Based on the visual and audio analyses performed: The video frames do not exhibit common AI-generated artifacts. The audio waveform and spectrum analysis do not show signs of synthetic generation. While these analyses suggest that the video and audio might be real, it is important to note that highly sophisticated AI-generated content can sometimes pass these basic checks. If further confirmation is needed, more specialized tools and forensic techniques should be employed.	Real	Real
Prompt 6: Tell me if synthesis artifacts are in the visual frames or audio. Must return with 1) yes or no only; 2) if yes, explain where the artifacts exist by answering in [region, artifacts] form	1. Yes • Visual: [All frames, High edge density] • Audio: [entire audio, Clipping detected], [entire audio, Silence detected]	Fake	Real
Prompt 7: I want you to work as a multimedia forensic expert for audio-visual deepfake detection. Check if the video has the artifact attribute listed in the following list and ONLY return the attribute number in this video. The artifact list is [1- unnatural face edges; 2- inconsistent lighting and shading; 3- irregular teeth shape or texture; 4-irregular lip movement; 5-inconsistent skin texture; 6- Spectral Artifacts; 7- Phoneme Artifacts; 8- Inconsistencies in Speech Patterns; 9- Voice Quality Issues; 10- lack of synchronization between audio and video]	The video exhibits the artifact corresponding to attribute number 2 - inconsistent lighting and shading	Fake	Fake

Figure 4: Demonstration of ChatGPT responses, which takes video and text prompts as input and produces audiovisual analysis, including explanations and authenticity predictions.

indicating that it is less effective in guiding the model to make accurate predictions. Its simplicity and lack of necessary information resulted in a rejection rate as high as 7.50%, preventing the model from making predictions for every video input. The numbers of TP, FP, TN, and FN are 7, 5, 13, and 12 respectively. In the deepfake video detection task, a higher TP value is desirable. However, instead of obtaining higher TP, P1 produced more TN, resulting in a lower recall of 0.368.

- Prompt P2: Similar to P1, P2 generates a binary response as to whether the video is real or not. Its lack of contextual information resulted in an even lower accuracy of 53.8%, but it was better than P1 in terms of rejection rate, which was 2.50%. The numbers of TP, FP, TN, and FN are 5, 3, 16, and 15, respectively. In terms of TP, P2 performed worse than P1. Due to the lower number of FP, its precision was slightly better than P1 at 0.625, while P2 had a higher number of FN, resulting in a lower recall than P1 at 0.250.

- Prompt P3: Unlike the binary response output in P1 and P2, P3 requires the model to return a probabilistic output, which results in an accuracy of 55.0% and a rejection rate of 0%, slightly better performance than P1 and P2. The numbers of TP, FP, TN, and FN are 5, 3, 17, and 15, respectively. Compared to P2, TN increases by 1. The precision and recall rates are 0.625 and 0.250, the same as P2.

- Prompt P4: P4 has improved accuracy compared to the previous three prompts. This prompt contains the term "AI-generated" and requires a probability score, resulting in better performance and increased accuracy to 57.5%. The numbers of TP, FP, TN, and FN are 12, 9, 11, and 8, respectively. The precision rate is 0.571, which is slightly lower due to the higher number of FP compared to the previous three prompts. However, due to the lower number of FN, P4 achieves a better recall rate of 0.600.

- Prompt P5: Like P1, P2, and P3, P5 appeared to be less effective due to its lack of specificity and manipulation details. P5 achieved an accuracy of 52.5%. The numbers of TP, FP, TN, and FN are 4, 3, 17, and 16, respectively. P5 has the same precision as P4, but its recall is poor at 0.200. The main reason for the extremely low recall rate is the small number of TP, only 4 out of 20 fake samples were correctly detected as fakes.

- Prompt P6: With an accuracy of 65.0%, higher precision and recall, and a rejection rate of 0%, P6 is by far the best performing prompt. Unlike the previous prompts, P6 focuses on visual and acoustic artifacts present in visual and audio modalities, allowing the underlying multimodal model to make the final prediction/decision more effectively. The numbers of TP, FP, TN, and FN are 17, 11, 9, and 3, respectively. P6 has a higher number of TP and TN than the previous prompts.

- Prompt P7: Compared to P6, the contextual details in P7 text prompt narrow the focus of the LLM model to specific artifacts and manipulations in both modalities, yielding more accurate and reliable results. P7 achieved an accuracy of 65.0%, a rejection rate of 0%, and the highest recall among all text prompts due to the larger number of TP. The numbers of TP, FP, TN, and FN are 18, 12, 8, and 2, respectively.

In summary, prompts based on simple binary responses often lack the necessary clarity and details to effectively leverage the multimodal knowledge of the LLM. Therefore, prompts like P1 to P5 resulted in lower accuracy. In contrast, more context-rich, artifact-based, and detailed-oriented prompts, such as P6 and P7, outperformed other simpler prompts. These more effective prompts leverage the multimodal capabilities and underlying knowledge of the LLM, yielding detection results aware of specific artifacts and manipulations in both modalities.

4.3.2 Comparing Different Text Prompts on the DFDC Test Set

Table 2 lists the performance of different text prompts (P1-P7 in Section 3.1) evaluated on the DFDC test set. Figure 5 shows a bar graph comparison of TP, TN, FP, and FN (see Section 4.2) for different text prompts.

Table 2: Precision, Recall, F1 Score, Rejection Rate, and Accuracy for different text prompts (P1P7) on the DFDC test set.

Prompt	Precision	Recall	F1 Score	Rejection Rate	Accuracy
P1	0.481	0.684	0.565	2.50	48.72
P2	0.435	0.500	0.465	0.00	42.50
P3	0.333	0.200	0.250	2.50	38.46
P4	0.483	0.700	0.571	0.00	47.50
P5	0.500	0.684	0.578	5.00	50.00
P6	0.500	1.00	0.667	0.00	50.00
P7	0.512	1.00	0.678	0.00	52.50

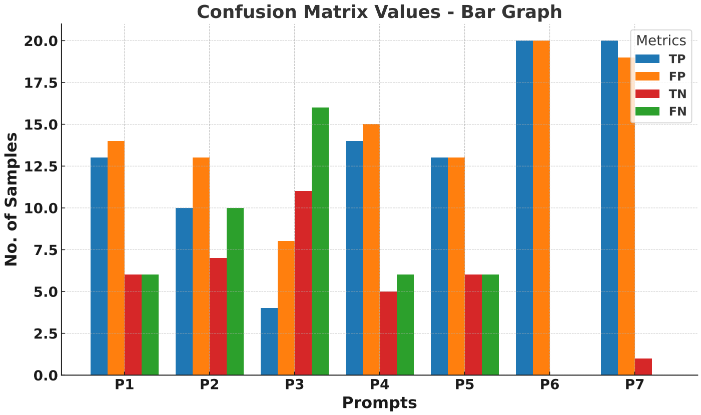


Figure 5: Bar graph comparing the number of True Positives, False Positives, True Negatives, and False Negatives for different text prompts on the DFDC test set.

From Table 2, we again see that context-rich, artifact-based, and detail-focused prompt P7 outperforms the other prompts. However, comparing Table 2 with Table 1 and Figure 5 with Figure 3, it is clear that the DFDC test set is more challenging than the FakeAVCeleb test set. The DFDC test set poses a significant challenge for deepfake detection due to its dominant visual manipulation and diverse real-world conditions, including illumination variation, occlusion, side-posed faces, multiple speakers, and background noise interference. Unlike the FakeAVCeleb test set, where videos typically feature a single, front-facing speaker under controlled conditions, the complexity of the DFDC test videos reduces the effectiveness of prompt-based detection using ChatGPT. The poor performance on the DFDC test set highlights the limitations of language-model-driven approaches in this setting. These findings suggest that datasets with high visual and acoustic variation require more sophisticated, artifact-aware tools to reliably detect forgeries.

4.3.3 Comparing ChatGPT with Human and AI Models

The detection performance of human evaluators and various state-of-the-art deep learning-based models on the FakeAVCeleb and DFDC test sets was compared in [20], as summarized in Table 3. In the human subjective test, each subject evaluated the same set of videos twice, in a different playback order. Phase 1 in the table represents the average accuracy of all testers in the first round, and Phase 2 represents the average accuracy of the second round. The results showed that humans performed better in the second round, but the difference between the two rounds was not significant. The average accuracy of human evaluators was 65.64% on the FakeAVCeleb test set and 63.84% on the DFDC test set, which serve as the baseline for comparison in our study. Among AI models, Lipforensics [17] focuses on semantic inconsistencies in the mouth region, achieving an accuracy of 92.50% on the FakeAVCeleb test set and 77.50% on the DFDC test set. The AV-Lip-Sync model [42] exploits synchronization between audiovisual modalities and achieves 87.50% and 75.00% accuracy on the two test sets, respectively, which is slightly lower than the accuracy of Lipforensics. The remaining three models AV-Lip-Sync+ [43], CNN-Ensemble [18], and AVTENet [19] all achieve a higher accuracy of 97.50% on the FakeAVCeleb test set, and 80.00%, 72.50%, and 80.00% on the DFDC test set, respectively. Note that all 5 AI models are trained on the FakeAVCeleb dataset. Therefore, these models perform significantly worse when tested on the DFDC test set due to the domain shift between the two datasets and the greater diversity and less controlled conditions in the DFDC test set. This result highlights the challenge of generalizing AI models to datasets with different distribution characteristics. The comparison results in Table 3 show that ChatGPT performs on par with humans when provided with appropriate prompts on the FakeAVCeleb test set, and slightly worse than humans on the more difficult DFDC test set, but both perform much worse than today's AI detection models. The higher accuracy of these deepfake detection models is attributed to their training on the multimodal FakeAVCeleb dataset.

Table 3: Comparison of detection accuracy between humans, AI models, and ChatGPT on the FakeAVCeleb and DFDC test sets.

Category	**Method**	**FakeAVCeleb**	**DFDC**
Human	Human (Phase I) [20]	63.30	59.52
	Human (Phase II) [20]	67.98	68.16
	Human (Overall) [20]	65.64	63.84
AI Models	LipForensics [17]	92.50	77.50
	AV-Lip-Sync [42]	87.50	75.00
	AV-Lip-Sync+ [43]	97.50	80.00
	CNN-Ensemble [18]	97.50	72.50
	AVTENet [19]	97.50	80.00
ChatGPT	P1	54.05	48.72
	P2	53.85	42.50
	P3	55.00	38.46
	P4	57.50	47.50
	P5	52.50	50.00
	P6	65.00	50.00
	P7	65.00	52.50

5 Ablation Study

5.1 Effectiveness of Prompts

Initially, we tested some basic prompts by mentioning only “video” and no mention of “audio”, and executed custom prompts sequentially within one session. The following are the custom text prompts:

- Tell me if this is an AI-generated video. Answer yes or no.
- Tell me if this is a real video. Answer yes or no.
- Tell me the probability of this video being real. Answer a probability score between 0 and 100.
- Tell me the probability of this video being AI-generated. Answer a probability score between 0 and 100.
- Tell me if this is a real or AI-generated video.
- 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.
- I want you to work as a video forensic expert for AI-generated faces. Check if the video has the artifact attribute in the following list and

> ONLY return the attribute number in this image. The artifact list is [1-asymmetric eye iris; 2-irregular 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-unnatural edges; 10-lack of synchronization between audio and video].

As can be seen from Table 4, the rejection rate is higher and the accuracy is lower, compared with the results in Table 1. Based on these results, we made several observations. First, only mentioning the video in the prompt causes the model to focus mainly on visual frames without analyzing the audio modality. To obtain the desirable output from an LLM-based model, prompts need to be specific and context-rich. Second, when prompts are fed sequentially, the model takes into account the context of the results of previous prompts, affecting its response to the current prompt. To obtain independent and unbiased results from prompts, we must feed the input video and prompt independently within a session to eliminate the effects of contextual bias from previous prompts. Third, our experiments show that prompts must contain terms relevant to acoustic analysis in order for the model to effectively analyze the audiovisual content in a given video.

Table 4: Precision, Recall, F1 Score, Rejection Rate, and Accuracy for different video-only mention prompts (P1-P7) on the FakeAVCeleb test set.

Prompt	Precision	Recall	F1 Score	Rejection Rate	Accuracy
P1	0.545	0.500	0.522	52.5	42.11
P2	0.467	0.500	0.483	40.0	37.50
P3	0.542	0.813	0.650	27.5	51.72
P4	0.480	0.750	0.585	17.5	48.48
P5	0.545	0.800	0.649	22.5	56.66
P6	0.600	0.474	0.529	2.50	57.89
P7	0.333	0.167	0.222	10.0	41.67

5.2 Failure Case Study

The reasons for detection failure vary depending on both the input prompt and video content. Through our experiments and careful analysis, we observed several factors that lead to inaccurate decisions in the multimodal ChatGPT model.

One factor is a high silence ratio in the speech content, which may indicate robotic/synthetic audio since the speech generation pipeline excludes environmental noise. However, videos with clean/enhanced speech do not always indicate synthesis or voice spoofing. Conversely, adding synthetic environmental noise to clean audio can mislead the model, leading to inaccurate predictions. The high silence rate combined with unnatural pauses in the acoustic modality can lead to an increased number of false positives in the model.

While unimodal/multimodal deep features and audiovisual correlation features are effective in various multimodal tasks, ChatGPT mainly relies on hand-crafted features and traditional functions in computer vision and speech processing libraries, including OpenCV, librosa, numpy, wav, and skimage, for visual and acoustic analysis. Furthermore, existing deep learning-based pretrained foundation models [44, 13, 3, 15] and frameworks (such as Tensorflow or Pytorch) are not used to analyze video content for possible artifacts and forgeries. These two shortcomings limit the ChatGPT method from effectively analyzing video content and result in poor performance compared to state-of-the-art forensic models.

In the context of audiovisual video forgery detection, if any modality (audio or video) is fake, the final prediction should be classified as fake. However, we observed that the overall probability score, calculated as the average of the audio and video scores, can lead the model to make incorrect predictions. If the score of one of the modalities dominates, the final prediction tends to reflect that modality, compromising the overall accuracy and classification results.

6 Limitations and Discussion

Although LLM-based models are superior to end-end learning-based black box models in terms of generalization, interpretability, and intuitive user interface for end users, they still have limitations. LLM-based models require domain knowledge for multimedia forensics tasks to design more effective prompts to exploit their underlying multimodal capabilities. Simple binary prompts are ineffective and yield lower accuracy and higher rejection rates. Furthermore, ChatGPT uses traditional techniques to analyze forgery in audiovisual content and has no access to pretrained models specifically trained for multimodal forgery detection tasks, resulting in lower accuracy even when the text prompts are effective and contextually rich. Given these limitations, the multimedia forensics community must focus on cutting-edge, end-to-end learning-based techniques to develop more robust, generalizable, and explainable audiovisual deepfake detectors.

To enhance ChatGPTs performance on audiovisual deepfake detection, we believe it is essential to go beyond traditional rule-based methods, which are often ineffective for complex manipulations. One way to improve performance is to integrate state-of-the-art pretrained foundation models, such as AV-HuBERT or Vision Transformers that provide rich audiovisual representations. These models can provide intermediate features or insights that

ChatGPT can interpret and reason about more effectively. Additionally, fusing specialized deepfake detectors (e.g., models targeting lip-sync errors, voice inconsistencies, or facial reenactments) can provide complementary signals. ChatGPT can then be guided to integrate these outputs to form a more robust final prediction. In addition, carefully crafted prompts containing keywords describing temporal or multimodal inconsistencies can also help the model reason more like a forensic expert. Further improvements can be achieved by tuning ChatGPT using domain-specific data, enabling it to better understand subtle artifacts common in manipulated media. Lastly, incorporating uncertainty estimates from supporting models and introducing human feedback where needed can help make the system more reliable and explainable.

7 Conclusions

In this study, we investigated the detection capabilities of a large language model (LLM) (i.e., ChatGPT) in the multimodal forgery detection task. We compared its performance with that of end-to-end multimedia forensic methods and human capabilities. Our results showed that, although ChatGPT was not explicitly designed for multimedia forgery detection tasks, its performance was comparable to human detection performance, demonstrating its potential in this field. A notable advantage of using LLMs in video forensics is their ability to generalize effectively because these models are learned from a wide range of datasets, unlike end-to-end models that are typically learned from specific video deepfake datasets. Additionally, LLMs provide superior interpretability compared to deep learning-based forensic methods, which, while capable of identifying specific visual and acoustic artifacts, typically serve as black-box models with limited interpretability. In future work, we aim to combine LLM-based models with deep learning-based forensic models to enhance interpretability and further contribute more interpretable and transparent deepfake detection tools to the forensics community.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: a compact facial video forgery detection network”, in *Proceedings of the IEEE International Workshop on Information Forensics and Security*, 2018, 1–7.

[2] Y. Apolo and K. Michael, “Beyond A Reasonable Doubt? Audiovisual Evidence, AI Manipulation, Deepfakes, and the Law”, *IEEE Transactions on Technology and Society*, 5(2), 2024, 156–68.

[3] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lui, and C. Schmid, “ViViT: A Video Vision Transformer”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021, 6836–46.

[4] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, *et al.*, “A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity”, in *Proceedings of the International Joint Conference on Natural Language Processing and Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, 2023, 675–718.

[5] S. S. Biswas, “Potential use of Chat GPT in global warming”, *Annals of Biomedical Engineering*, 51(6), 2023, 1126–7.

[6] S. S. Biswas, “Role of Chat GPT in public health”, *Annals of Biomedical Engineering*, 51(5), 2023, 868–9.

[7] K. M. Caramancion, “Harnessing the power of ChatGPT to decimate mis/disinformation: Using ChatGPT for fake news detection”, in *Proceedings of the IEEE World AI IoT Congress*, 2023, 42–46.

[8] J. R. Carvalko, “Generative AI, Ingenuity, and Law”, *IEEE Transactions on Technology and Society*, 5(2), 2024, 169–82.

[9] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang, and L. Nie, “Voice-face homogeneity tells deepfake”, *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3), 2023, 1–22.

[10] A. Chintha, B. Thai, S. J. Sohrawardi, K. Bhatt, A. Hickerson, M. Wright, and R. Ptucha, “Recurrent convolutional structures for audio spoof and video deepfake detection”, *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 2020, 1024–37.

[11] I. DeAndres-Tame, R. Tolosana, R. Vera-Rodriguez, A. Morales, J. Fierrez, and J. Ortega-Garcia, “How Good is ChatGPT at Face Biometrics? A First Look into Recognition, Soft Biometrics, and Explainability”, *IEEE Access*, 12, 2024, 34390–401.

[12] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, “The DeepFake Detection Challenge (DFDC) Dataset”, *arXiv preprint arXiv:2006.07397*, 2020.

[13] A. Dosovitskiy, “An Image is Worth 16x16 Words: Transformers for image recognition at scale”, *arXiv preprint arXiv:2010.11929*, 2020.

[14] Á. Figueira and L. Oliveira, “The current state of fake news: challenges and opportunities”, *Procedia Computer Science*, 121, 2017, 817–25.

[15] Y. Gong, Y.-A. Chung, and J. Glass, “AST: Audio Spectrogram Transformer”, in *Proceedings of the Interspeech Conference*, 2021, 571–5.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks”, *Communications of the ACM*, 63(11), 2020, 139–44.

[17] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, “Lips Don’t Lie: A Generalisable and Robust Approach To Face Forgery Detection”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, 5039–49.

[18] A. Hashmi, S. A. Shahzad, W. Ahmad, C. W. Lin, Y. Tsao, and H.-M. Wang, “Multimodal forgery detection using ensemble learning”, in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2022, 1524–32.

[19] A. Hashmi, S. A. Shahzad, C.-W. Lin, Y. Tsao, and H.-M. Wang, “AVTENet: Audio-visual transformer-based ensemble network exploiting multiple experts for video deepfake detection”, *arXiv preprint arXiv:2310.13103*, 2023.

[20] A. Hashmi, S. A. Shahzad, C.-W. Lin, Y. Tsao, and H.-M. Wang, “Unmasking Illusions: Understanding Human Perception of Audiovisual Deepfakes”, *arXiv preprint arXiv:2405.04097*, 2024.

[21] J. He, L. Li, W. Yao, and H. Gao, “Exploring Future Education: The Innovative Integration and Practice of Multimodal Learning and ChatGPT”, in *Proceedings of the International Conference on Computer Science, Engineering, and Education*, 2024, 18–23.

[22] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models”, in *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 33, 2020, 6840–51.

[23] H. Ilyas, A. Javed, and K. M. Malik, “AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio-visual deepfakes detection”, *Applied Soft Computing*, 136, 2023, 110124.

[24] S. Jia, R. Lyu, K. Zhao, Y. Chen, Z. Yan, Y. Ju, C. Hu, X. Li, B. Wu, and S. Lyu, “Can ChatGPT detect deepfakes? a study of using multimodal large language models for media forensics”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, 4324–33.

[25] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu, *et al.*, “Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis”, in *Proceedings of the Advances in Neural Information Processing Systems*, Vol. 31, 2018, 4485–95.

[26] B. Kaddar, S. A. Fezza, Z. Akhtar, W. Hamidouche, A. Hadid, and J. Serra-Sagristà, “Deepfake detection using spatiotemporal transformer”, *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(11), 2024, 1–21.

[27] S. Karnouskos, “Artificial intelligence in digital media: The era of deepfakes”, *IEEE Transactions on Technology and Society*, 1(3), 2020, 138–47.

[28] H. Khalid, M. Kim, S. Tariq, and S. S. Woo, “Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors”, in *Proceedings of the Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection*, 2021, 7–15.

[29] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, “FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset”, in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[30] J. Koco, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydo, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, *et al.*, “ChatGPT: Jack of all trades, master of none”, *Information Fusion*, 99, 2023, 101861.

[31] I. Korshunova, W. Shi, J. Dambre, and L. Theis, “Fast Face-swap Using Convolutional Neural Networks”, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, 3677–85.

[32] A. O. Kwok and S. G. Koh, “Deepfake: a social construction of technology perspective”, *Current Issues in Tourism*, 24(13), 2021, 1798–802.

[33] X. Liao, K. Li, X. Zhu, and K. R. Liu, “Robust detection of image operator chain with two-stream convolutional neural network”, *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 2020, 955–68.

[34] K. Lutz and R. Bassett, “Deepfake Detection with Inconsistent Head Poses: Reproducibility and Analysis”, *arXiv preprint arXiv: 2108.12715*, 2021.

[35] O. Mayer and M. C. Stamm, “Exposing fake images with forensic similarity graphs”, *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 2020, 1049–64.

[36] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez, “GANprintR: Improved fakes and evaluation of the state of the art in face manipulation detection”, *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 2020, 1038–48.

[37] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-Forensics: Using capsule networks to detect forged images and videos”, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, 2307–11.

[38] Y. Nirkin, Y. Keller, and T. Hassner, “FSGAN: Subject Agnostic Face Swapping and Reenactment”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, 7184–93.

[39] T. B. Patel and H. A. Patil, “Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech.”, in *Proceedings of the Interspeech Conference*, 2015, 2062–6.

[40] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild", in *Proceedings of the ACM International Conference on Multimedia*, 2020, 484–92.

[41] A. Ray, "Disinformation, deepfakes and democracies: The need for legislative reform", *The University of New South Wales Law Journal*, 44(3), 2021, 983–1013.

[42] S. A. Shahzad, A. Hashmi, S. Khan, Y.-T. Peng, Y. Tsao, and H.-M. Wang, "Lip sync matters: A novel multimodal forgery detector", in *Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2022, 1885–92.

[43] S. A. Shahzad, A. Hashmi, Y.-T. Peng, Y. Tsao, and H.-M. Wang, "AV-Lip-Sync+: Leveraging AV-HuBERT to exploit multimodal inconsistency for video deepfake detection", *arXiv preprint arXiv: 2311.02733*, 2023.

[44] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction", in *Proceedings of the International Conference on Learning Representations*, 2021.

[45] M. Todisco, H. Delgado, and N. Evans, "A New Feature for Automatic Speaker Verification Anti-Spoofing: Constant Q Cepstral Coefficients", in *Proceedings of the Speaker and Language Recognition Workshop (Odyssey)*, Vol. 45, 2016, 283.

[46] C. Vaccari and A. Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news", *Social Media+ Society*, 6(1), 2020.

[47] T. Vaikunta Pai, P. Nethravathi, R. Birau, V. Popescu, B. Karthik Pai, and P. V. Naik, "Multimodal ChatGPT: Extending ChatGPT to enable rich multimodal conversations using deep neural network", *Journal of Intelligent & Fuzzy Systems*, 2024, 1–17.

[48] L. Verdoliva, "Media forensics and deepfakes: an overview", *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 2020, 910–32.

[49] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech.", in *Proceedings of the Interspeech Conference*, 2015, 2092–6.

[50] G. Wu, W. Wu, X. Liu, K. Xu, T. Wan, and W. Wang, "Cheap-fake Detection with LLM using Prompt Engineering", in *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops*, 2023, 105–9.

[51] H. Wu, H.-C. Kuo, N. Zheng, K.-H. Hung, H.-Y. Lee, Y. Tsao, H.-M. Wang, and H. Meng, "Partially fake audio detection by self-attention-based fake span discovery", in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, 9236–40.

[52] Y. Xu, P. Terhöst, M. Pedersen, and K. Raja, “Analyzing Fairness in Deepfake Detection With Massively Annotated Databases”, *IEEE Transactions on Technology and Society*, 5(1), 2024, 93–106.

[53] Z. Yan, K. Zhang, R. Zhou, L. He, X. Li, and L. Sun, “Multimodal ChatGPT for medical applications: an experimental study of GPT-4V”, *arXiv preprint arXiv:2310.19061*, 2023.

[54] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren, “AvoiD-DF: Audio-visual joint learning for detecting deepfake”, *IEEE Transactions on Information Forensics and Security*, 18, 2023, 2015–29.

[55] X. Yang and J. Zhou, “Research about the Ability of LLM in the Tamper-Detection Area”, *arXiv preprint arXiv:2401.13504*, 2024.

[56] Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, Z. Liu, C. Liu, M. Zeng, and L. Wang, “MM-REACT: Prompting ChatGPT for Multimodal Reasoning and Action”, *arXiv preprint arXiv:2303.11381*, 2023.

[57] G. P. Zachary, “Digital manipulation and the future of electoral democracy in the US”, *IEEE Transactions on Technology and Society*, 1(2), 2020, 104–12.

[58] D. Zhang, F. Lin, Y. Hua, P. Wang, D. Zeng, and S. Ge, “Deepfake video detection with spatiotemporal dropout transformer”, in *Proceedings of the ACM International Conference on Multimedia*, 2022, 5833–41.

[59] Y. Zhou and S.-N. Lim, “Joint audio-visual deepfake detection”, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 14800–9.

[60] H. Zou, M. Shen, Y. Hu, C. Chen, E. S. Chng, and D. Rajan, “Cross-modality and within-modality regularization for audio-visual deepfake detection”, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2024, 4900–4.