國立臺灣大學電機資訊學院資訊網路與多媒體研究所

碩士論文

Graduate Institute of Networking and Multimedia

College of Electrical Engineering and Computer Science

National Taiwan University

Master's Thesis

# 多模態大型語言模型之可解釋深度偽造檢測基準

# MMIDBench: Multimodal Interpretable Deepfake Detection Benchmark for Multimodal Large Language Models

黃康洋

Kang-Yang Huang

指導教授: 鄭文皇 博士

Advisor: Wen-Huang Cheng Ph.D.

中華民國 114 年 7 月

July, 2025

doi:10.6342/NTU202503386

# 國立臺灣大學碩士學位論文
# 口試委員會審定書

**MASTER'S THESIS ACCEPTANCE CERTIFICATE**
**NATIONAL TAIWAN UNIVERSITY**

## 多模態大型語言模型之可解釋深度偽造檢測基準

## MMIDBench: Multimodal Interpretable Deepfake Detection Benchmark for Multimodal Large Language Models

本論文係 黃康洋 （學號 R11944070）在國立臺灣大學資訊網路與多媒體研究所完成之碩士學位論文，於民國 114 年 7 月 30 日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Graduate Institute of Networking and Multimedia on 30 July 2025 have examined a Master's Thesis entitled above presented by HUANG, KANG-YANG (student ID: R11944070) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

鄭文皇　　莊永裕

（指導教授 Advisor）

黃敬群

系（所）主管 Director: 鄭卜壬

# Acknowledgements

Throughout my master's studies, I have been profoundly grateful for the unwavering guidance and encouragement of my advisor, Professor Wen-Huang Cheng, and Professor Hong-Han Shuai. It has been a true privilege to work under their supervision. Their expertise and insightful advice have been invaluable in shaping and completing this thesis.

I am sincerely grateful to Dr. Jhih-Ciang Wu and Dr. Jordi Sanchez-Riera for their ongoing guidance, which greatly enhanced the productivity and efficiency of my research.

I also thank Dr. Ling Lo for her valuable insights and for highlighting key issues and perspectives that helped me focus on the most important aspects of my work.

I am also deeply appreciative of my lab mates, particularly Ling Zou, Jian-Yu Jiang-Lin, Yu-Wen Tseng, Chu-Jun Peng, Feng-Kai Huang, Kun-Hsiang Lin, Shang-Fu Chen, Tai-Ming Huang, Yi Yao, Si-Yu Lu, Tai-i Chen, Sheng-Ping Yang, Xingyi Zheng, TeI Hung, Rong Chao, Ruyi Xu, Ching Chia Huang, and Jia-Wei Liao. Their support and friendship have made my research journey both enjoyable and rewarding.

Ultimately, I am deeply grateful to my family and my girlfriend, Ling Zou, for their persistent love and support, which made the completion of this work possible.

This research was funded by NVIDIA and utilized NVIDIA A100 GPUs.



doi:10.6342/NTU202503386

# 摘要

生成式人工智慧透過運用多樣化的輸入條件，徹底革新了多媒體內容的創作方式。然而，隨著這些模型日益進步，檢測由人工智慧生成的內容，特別是深度偽造（DeepFake），變得愈發困難。對深偽技術日益增加的關注，使得檢測方法，尤其是多模態大型語言模型在辨識深偽內容方面的效能，成為研究重點。多模態大型語言模型不僅能透過提供決策解釋來提升深偽檢測的透明度，區分真實與合成內容的過程同時也是對其感知與推理能力的嚴格考驗。

為了應對這些挑戰，我們提出了 MMIDBench，一個精心設計、全面評估多模態大型語言模型能力的多模態基準。MMIDBench 涵蓋多種最先進的深偽生成模型，橫跨影像、影片與音訊，包含 6 種不同的深偽任務。該基準包含 10k 道題目，涵蓋二元選擇、多選題及開放式問答等多種題型，能夠對多模態大型語言模型進行深入評估。我們利用 MMIDBench 評測了 5 款閉源多模態大型語言模型，揭示了它們在深偽檢測上的優勢與現階段的侷限。

**關鍵字：**多模態大型語言模型、可解釋性、深度偽造檢測、深度偽造生成、影像編輯、語音合成、擴散模型、生成式人工智慧


doi:10.6342/NTU202503386

# Abstract

Generative artificial intelligence has revolutionized how multimedia content is created by utilizing diverse input conditions. However, as these models become more advanced, detecting AI-generated content, particularly DeepFakes, has grown increasingly challenging. Rising concerns over DeepFakes have heightened interest in detection methods, specifically the effectiveness of multimodal large language models (MLLMs) in identifying them. MLLMs not only improve the transparency of DeepFake detection by providing explanations for their decisions, but the process of distinguishing authentic from synthetic content also serves as a robust test of their perceptual and reasoning skills.

To address these challenges, we introduce MMIDBench, a comprehensive multimodal benchmark meticulously crafted to assess the capabilities of MLLMs. MMIDBench features a variety of state-of-the-art DeepFake generative models spanning images, videos, and audio, encompassing 6 distinct DeepFake tasks. The benchmark comprises 10k questions, including binary, multiple-choice, and open-ended formats, enabling an in-depth


doi:10.6342/NTU202503386

assessment of MLLMs. We evaluated 5 proprietary MLLMs with MMIDBench, revealing both their strengths and current limitations in DeepFake detection.

**Keywords:** Multimodal Large Language Models, Interpretability, Deepfake Detection, Deepfake Generation, Image Editing, Voice Synthesis, Diffusion Models, Generative Artificial Intelligence



doi:10.6342/NTU202503386

# Contents





doi:10.6342/NTU202503386

doi:10.6342/NTU202503386

# List of Figures





doi:10.6342/NTU202503386

doi:10.6342/NTU202503386

# List of Tables




doi:10.6342/NTU202503386

doi:10.6342/NTU202503386

# Chapter 1 Introduction

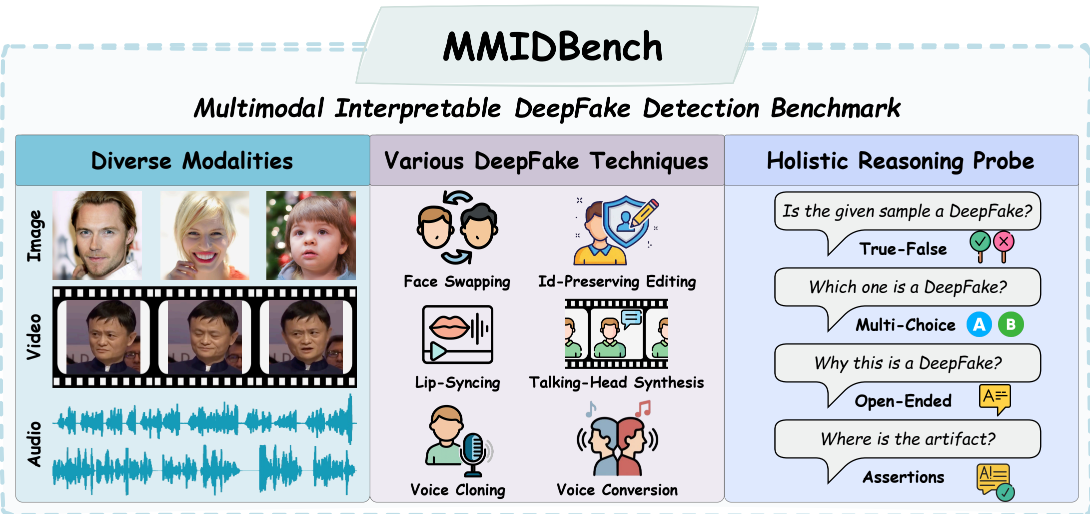


Figure 1.1: **MMIDBench** features three primary attributes that set it apart. First, it encompasses a wide range of modalities, including images, videos, and audio. Second, it incorporates an array of DeepFake methods, such as face-swapping, identity-preserving editing, lip-syncing, audio-driven talking-head synthesis, voice cloning, and voice conversion. Third, it integrates the Holistic Reasoning Probe, which delivers a coarse-to-fine assessment of the reasoning and perceptual capabilities of Multimodal Large Language Models.

The rapid advancement of generative artificial intelligence (AI) has transformed the landscape of multimedia content creation, enabling the synthesis of highly realistic images, videos, and audio from diverse inputs. This progress, while innovative, has amplified the risks associated with DeepFakes, which are AI-generated forgeries designed to mimic authentic media. Such manipulations can facilitate the spread of disinformation, erode public trust, and inflict reputational harm, underscoring the urgent need for effective detection mechanisms. Traditional approaches to DeepFake detection often concentrate


doi:10.6342/NTU202503386

on isolated modalities or specific forgery techniques, rendering them inadequate against the sophisticated, multimodal threats emerging today [82]. In response, there is growing interest in leveraging multimodal large language models (MLLMs) not only for their detection prowess but also for their ability to provide interpretable explanations, enhancing transparency and reliability in identifying synthetic content [60, 115].

At the heart of this challenge lies the demand for benchmarks that rigorously evaluate MLLMs' perceptual and reasoning capabilities in a DeepFake context. MLLMs offer a promising avenue for interpretable detection, as they can articulate the rationale behind their judgments, bridging the gap between automated analysis and human-understandable insights. However, existing evaluation frameworks frequently overlook the multifaceted nature of DeepFakes, failing to encompass diverse modalities and generation methods [56, 105, 124]. To bridge this gap, we introduce MMIDBench, a comprehensive benchmark tailored for assessing MLLMs in interpretable DeepFake detection. Our key contributions include dramatically increasing forgery diversity by incorporating content from approximately 6distinct DeepFake techniques—such as face-swapping, identity-preserving editing, lip-syncing, talking-head synthesis, voice cloning, and voice conversion—far surpassing the limited 1 to 4 types in prior datasets by providing a unified, multimodal framework spanning images, videos, and audio, with multi-task evaluation across authenticity verification, explainability through reasoning or description, and perception of manipulated regions, and delivering a large-scale, balanced dataset of 10kquestions in formats like binary, multiple-choice, and open-ended, complete with matched real-fake pairs and strict data hygiene protocols to prevent leakage. These innovations address critical limitations in benchmarks like MMTD-Set [115], SID-Set [40], and LOKI [124], fostering the development of robust, generalizable, and interpretable detection systems.


doi:10.6342/NTU202503386

As illustrated in Figure 1.1, MMIDBench stands out due to three key attributes that together create a comprehensive evaluation framework. First, it adopts a multimodal approach, encompassing images, videos, and audio. This ensures that detection capabilities are tested across a wide range of media types commonly targeted by DeepFake technologies. By addressing the limitations of unimodal benchmarks, MMIDBench simulates real-world scenarios where forgeries often combine multiple forms of data.

Second, the benchmark includes a diverse set of DeepFake techniques. These range from face-swapping, which alters identities in visuals, to identity-preserving editing, which subtly modifies appearances without changing core features. It also covers lip-syncing, which aligns speech with video, audio-driven talking-head synthesis, which animates faces based on voice inputs, voice cloning, which replicates individual speech patterns, and voice conversion, which transforms audio characteristics. This variety reflects the growing sophistication of generative attacks and enables a detailed evaluation of model robustness.

Third, MMIDBench integrates the Holistic Reasoning Probe, a structured evaluation tool that assesses both high-level perceptual skills and fine-grained reasoning abilities. It examines tasks such as identifying visual inconsistencies and detecting manipulation traces through logical inference. Additionally, the probe enhances interpretability by producing explanatory outputs that clarify the model's decision-making process.

Building on these foundations, MMIDBench incorporates state-of-the-art DeepFake generative models across 6 distinct tasks, featuring 10k questions in varied formats like binary, multiple-choice, and open-ended. This design facilitates a thorough examination of MLLMs' effectiveness in distinguishing genuine from manipulated media, while em-


doi:10.6342/NTU202503386

phasizing interpretability through explanatory outputs. Our evaluation of 5 proprietary MLLMs, including models like GPT-4o [42] and Gemini-2.5-Pro [17], reveals critical insights into their performance, highlighting both advancements and persistent limitations in handling complex forgeries, with comparisons to human evaluators and specialized expert models. By addressing these aspects, MMIDBench not only advances the field of DeepFake detection but also contributes to a holistic defense system that safeguards data privacy, security, and trust across the AI lifecycle, from input protection to output verification.

In the following sections, we detail the construction of MMIDBench, present our experimental findings, and discuss implications for future research in interpretable AI-driven detection.


doi:10.6342/NTU202503386

# Chapter 2 Related Work

## 2.1 DeepFake Generation

DeepFake technology leverages artificial intelligence to generate or manipulate media, resulting in highly convincing but synthetic images, videos, and audio. The field has evolved rapidly, progressing from early, labor-intensive computer graphics techniques to modern deep learning systems capable of automatically producing photorealistic content. Initial breakthroughs were achieved with autoencoders, particularly Variational Autoencoders (VAEs) [86, 94], but the introduction of Generative Adversarial Networks (GANs) [29, 48, 49] marked a turning point, laying the groundwork for many current media manipulation methods. More recently, diffusion models [36, 81, 85, 88] and Neural Radiance Fields (NeRF) [70] have set new benchmarks for realism and control in synthetic media generation.

A variety of DeepFake creation techniques have emerged, each focusing on altering specific aspects of media to make the process more accessible and effective. Identity swapping [41, 55, 61, 97, 142] replaces a person's identity in an image, video, or audio file while preserving other features such as facial expressions or speech. Expression or emotion swapping [25, 80, 91, 110, 121] modifies the emotional tone or facial expressions of a subject without changing their core identity. Facial attribute manipulation [4, 39, 102,


doi:10.6342/NTU202503386

129, 139] targets specific features like age, gender, or hairstyle, enabling nuanced edits while maintaining overall identity. Advanced methods such as talking face synthesis [5, 51, 68, 113, 126] generate synchronized audio and video of a person speaking, guided by text or audio input for realistic alignment.

Recent generative models have also enabled text-to-image [10, 23, 78] and text-to-video [9, 66] synthesis, allowing users to create entirely new media from text prompts, specifying details such as facial features, body pose, and actions. Content editing or partial synthesis [27, 28, 64, 119, 131] enables modification of specific parts of existing media —such as altering certain video frames or audio segments—while preserving the overall continuity and identity. Some methods are domain-agnostic, applicable across various media types, while others are domain-specific.

The realism achieved by current DeepFake generation methods, driven by the transition from GANs to diffusion models and the integration of 3D priors, makes synthetic media increasingly difficult to distinguish from authentic content. However, this progress brings significant challenges. The widespread availability and sophistication of DeepFake technology raise serious social concerns, including the spread of misinformation, creation of non-consensual material, and fraud, all of which threaten public trust and the integrity of digital information. For further discussion, see [19, 77].

## 2.2 DeepFake Detection

The field of DeepFake detection has advanced rapidly alongside the rise of synthetic media. Early detection strategies relied on analyzing hand-crafted features and intrinsic image statistics—such as eye blinking, head pose, and visual inconsistencies, to spot ma-


doi:10.6342/NTU202503386

nipulated content [18, 71, 120, 122]. While these classical methods performed well on familiar datasets, their accuracy dropped significantly when confronted with more sophisticated DeepFakes or data from unfamiliar sources, largely due to limited dataset diversity and poor generalization.

With the advent of deep learning, researchers began framing DeepFake detection as a binary classification task, leveraging deep neural networks (DNNs) to improve performance, utilizing pre-trained CNN architectures like VGG, ResNet, DenseNet, and XceptionNet, which have shown improved results, especially when tested on data from different sources [22, 73, 82, 109]. Ensemble learning and multi-task strategies have also been explored to further boost detection accuracy and resilience.

A persistent challenge in this domain is transferability—the ability of a model trained on one dataset or manipulation technique to perform well on unseen data. As DeepFake generation tools become more advanced and widespread, ensuring that detection systems can adapt to new and diverse manipulations is crucial. To tackle this, researchers have incorporated frequency domain analysis, attention mechanisms, and vision transformer (ViT) models to better capture global image context [24, 31, 43, 50]. Data augmentation and the use of synthetic training data have also been adopted to enhance model generalization [13, 37, 72, 89, 103, 114, 117].

Robustness is equally important, especially in practical scenarios where DeepFake media may be compressed, post-processed, or subjected to adversarial attacks. Models that excel on clean, curated datasets often falter when faced with degraded or intentionally manipulated inputs. To address this, researchers have developed techniques resilient to compression, noise, and adversarial perturbations [6, 32, 38, 90, 108]. Approaches such


doi:10.6342/NTU202503386

as multi-stream architectures, attention distillation, and adversarial training have been used to strengthen models against both passive and active threats. Recent work has also evaluated model robustness using established adversarial attack methods like FGSM [30] and PGD [69], emphasizing the ongoing need to defend against increasingly sophisticated manipulations [45, 112].

Interpretability remains another major concern. Although deep learning models can achieve high detection rates, their internal decision-making processes are often opaque, which is problematic for forensic and trust-critical applications. Efforts to make these models more transparent have included noise-based analysis, artifact localization at the region or pixel level, and the extraction of manipulation traces using denoising networks [15, 35, 44, 99, 130]. However, common visualization tools like heatmaps often fall short in clearly distinguishing real from fake content, highlighting the need for more effective interpretability solutions. While DeepFake detection has made significant strides, ongoing research continues to focus on improving transferability, interpretability, and robustness to ensure these systems remain effective and trustworthy in real-world applications.

## 2.3 Multimodal Large Language Models

The rapid advancement of Large Language Models (LLMs) like GPT [75] and LLaMA [93] has fueled growing interest in Multimodal Large Language Models (MLLMs), which integrate visual and textual data for more comprehensive understanding and reasoning. By aligning visual features with language representations, MLLMs produce outputs that are both contextually rich and diverse.

Early models such as Flamingo [2] introduced cross-attention mechanisms to incor-


doi:10.6342/NTU202503386

porate visual information into language modeling. Subsequent innovations like BLIP-2 [54] and InstructBLIP [21] improved this integration through novel architectures and instruction-following data. LLaVA [59] further advanced the field by creating large-scale instruction-based datasets, enhancing multimodal task performance. Other notable models, including MiniGPT-4 [140], Qwen2.5-Omni [111], and mPLUG-Owl [123], have each contributed unique methods for strengthening multimodal alignment and reasoning.

In digital forensics, MLLMs are increasingly effective for detecting manipulated content. Models like ForgeryGPT [60] and VLForgery [33] leverage multimodal reasoning to identify deepfakes and image forgeries with high precision. FakeShield [115] further enhance interpretability and localization of altered regions. These developments underscore the growing importance of MLLMs in both research and real-world applications, setting new standards for multimodal analysis and content verification.

## 2.4 Synthetic Data Detection Benchmark

The rapid proliferation of AI-generated media, such as DeepFakes, has driven the development of increasingly sophisticated detection technologies. This progress has been closely tied to the evolution of benchmarks, which have grown larger, more diverse, and more complex over time. Early research focused on the basic task of distinguishing authentic from manipulated content—a straightforward binary classification problem. Initial benchmarks, while foundational, were often limited in diversity and scope.

For example, the FFHQ [48] dataset provided high-quality images generated by a single GAN-based technique, serving as an early standard. FakeSpotter [96] expanded on this by including images from four different DeepFake methods, challenging models to


doi:10.6342/NTU202503386

generalize across manipulations. However, both datasets were restricted to images and binary authenticity classification, lacking support for multi-modal or multi-task evaluation. CNNSpot [98] addressed the need for scale, offering a large dataset for training data-hungry CNN-based detectors, but still focused on a single forgery type and binary classification. In the audio domain, the ASVspoof [101] 2019 challenge dataset became a key benchmark, providing samples of two types of audio spoofing and establishing standards for anti-spoofing in speaker verification systems. Recent datasets like Genimage [141] and GenVideo [8] have further increased scale, offering massive collections for training binary classifiers. However, these remain single-modality and do not address more advanced tasks such as localization or explainability, keeping the field's focus on scaling classification rather than deepening analytical capabilities.

A significant advancement came with the shift from binary classification to forgery localization—identifying not just whether media is fake, but precisely where manipulations occur, often at the pixel level. This required more detailed annotations and led to the development of spatially-aware detection models. ForgeryNet [34] marked a turning point by supporting both authenticity classification and forgery localization, providing millions of images with detailed tampering masks. This dual-task design made it a comprehensive and widely adopted benchmark. Subsequent datasets like UniversalFake [74] and Fake2M [67] reinforced this approach, further establishing the dual-task framework as standard in image forensics. Localization has also been introduced in the audio domain. FakeMusicCaps [16] enables both authenticity and localization tasks, allowing researchers to pinpoint manipulated segments within audio clips. Despite these advances, most benchmarks still lack interpretability—they indicate whether and where manipulation occurs, but not why a model made its decision. This *black box* effect limits real-world adoption,


doi:10.6342/NTU202503386

especially in fields like journalism, law enforcement, and national security, where transparent, evidence-based explanations are essential.

The next logical step is explainability: providing clear, human-understandable reasons for detection decisions. Advances in Multimodal Large Language Models (MLLMs), which excel at reasoning and generating natural language explanations, have enabled this shift. New benchmarks now assess not only detection accuracy but also the clarity and reliability of model explanations. FakeBench [56] was the first benchmark to evaluate explainable detection by MLLMs, pairing fake images with detailed natural language descriptions of manipulation artifacts. Building on this, FakeClue [105] provides detailed annotations of specific forgery artifacts, pushing models toward deeper, more interpretable explanations. FakeChain [57] introduces Chain-of-Thought (CoT) prompting, breaking down the reasoning process step by step and emphasizing logical verification.

However, these early explainable benchmarks primarily focused on images, leaving other modalities underexplored. IVY-FAKE [133] addressed this by introducing a large-scale, multimodal dataset for explainable detection across images and videos, with richly annotated samples and detailed explanations. The most comprehensive benchmark to date, LOKI [124], evaluates MLLMs' ability to detect synthetic content and provide logical explanations across video, images, 3D data, text, and audio. LOKI's multi-level evaluation framework rigorously tests perception, knowledge, and reasoning skills. LOKI's results highlight both the promise and current limitations of MLLMs: while they show strong potential as explainable detectors, their performance varies across modalities, and logical reasoning remains a challenge, pointing to important directions for future research.

Alongside narrative explanations from ML models, another approach to explainabil-


doi:10.6342/NTU202503386

ity focuses on the discriminative models' ability to localize and characterize manipulations. Here, explainability is measured by how well a model can identify tampered areas, spot temporal inconsistencies, or uncover logical contradictions within content.

Benchmarks like SynthScars [47] and MMTD-Set [115] exemplify this strategy. SynthScars targets the detection and localization of subtle *synthetic scars* left by generative methods, while MMTD-Set emphasizes analyzing temporal dynamics in videos, as forgeries often reveal themselves through temporal disruptions. This temporal localization provides a compelling form of explanation by pinpointing the exact moments of tampering. SID-Set [40] addresses the challenge of detecting subtle manipulations common on social media. Its three-part framework integrates detection, localization (via mask prediction), and explanation (through textual criteria), with detailed annotations for both fully synthetic and subtly edited images. This demands highly precise localization and clear, verifiable explanations. The GIM [11] dataset represents a major advance in image manipulation detection and localization, offering over a million pairs of real and AI-manipulated images across diverse classes and manipulation types. Its large scale and diversity enable robust evaluation, and the accompanying GIMFormer framework achieves state-of-the-art performance.

These two approaches—narrative explanation and forensic localization—reflect a split in how explainability is defined. Generative MLLMs are well-suited for narrative explanations, while discriminative models like CNNs and ViTs excel at localization. Rather than being mutually exclusive, these methods are complementary. An ideal detection system would both identify where manipulation occurs and clearly explain why it is considered a forgery. However, most current benchmarks focus on only one aspect, leaving a gap in evaluating both forms of explainability together.


doi:10.6342/NTU202503386

# Chapter 3    MMIDBench

## 3.1    Overview of MMIDBench

The field of explainable DeepFake detection has rapidly diversified, giving rise to a wide range of benchmarks [40, 47, 56, 105, 115, 124, 141] with distinct objectives and methodologies. Research has moved beyond basic binary classification on single modalities to address more complex challenges, including multimodal analysis, precise localization, and advanced reasoning.

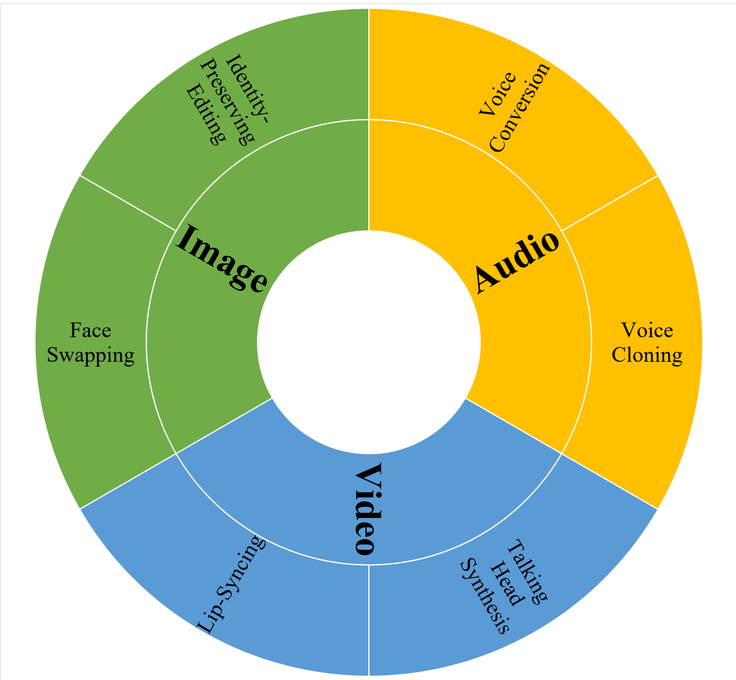


Figure 3.1: Statistical overview of MMIDBench, illustrating the distribution and detailed tasks across each modality, including audio, video, and image-based manipulations.


doi:10.6342/NTU202503386

As illustrated in Figure 3.1, MMIDBench encompasses a variety of modalities, including image, video, and audio, with over 6 specific techniques of DeepFake generation. The benchmark utilizes fine-grained anomaly annotations to construct a tiered variety of question types, including judgment questions, multiple-choice questions, and open-ended questions, totaling over 10k questions.

Table 3.1 outlines real-world DeepFake scenarios and the manipulation techniques applied across various data types, including image, video, and audio modalities. For instance, image-based techniques such as face swapping and identity-preserving editing, video-based methods like lip-syncing and talking-head syntheses, and audio-based approaches such as voice cloning and voice conversion are commonly exploited. These techniques enable misuse in scenarios ranging from identity theft and misinformation to impersonation and blackmail. In terms of breadth, MMIDBench covers a wider range of modalities, addressing diverse manipulation of content. In-depth, it extends beyond AI-generated content to include finer-grained DeepFake generation techniques, offering a comprehensive framework for evaluating and mitigating the risks associated with these advanced manipulations.

Table 3.1: Overview of DeepFake modalities, associated manipulation tasks, and representative real-world misuse scenarios.

| **Modality** | **Tasks** | **Real-world Scenarios** |
|---|---|---|
| Image | Face Swapping [41, 84, 138]<br>Identity-Preserving Editing [63, 92, 107] | Fake pornography, political smear campaigns, blackmail, identity theft, generating fake IDs, etc. |
| Video | Lip-Syncing [53, 135, 136]<br>Talking-Head Synthesis [20, 104, 132] | Fake pornography, fake video evidence in court, manipulating surveillance footage, impersonation in video calls, extortion, fake hostage videos, misinformation, etc. |
| Audio | Voice Cloning [26, 79]<br>Voice Conversion [3, 12, 62] | Impersonation in phone scams, bypassing voice authentication, manipulating audio evidence, harassment and defamation, blackmail, phishing, misinformation, , etc. |


doi:10.6342/NTU202503386

Table 3.2: A comprehensive comparison of prominent, publicly available datasets for DeepFake evaluation. For each dataset, we report the publication **Venue**, total dataset **Size**, the number of DeepFake generation methods included (**DeepFake Types**), **Data Modality**, and the types of ground-truth labels available: authenticity (**Au**), explainability (**Ex**), and manipulation localization (**Lo**). The **Real Pair** column (✓) indicates whether the dataset contains paired real and fake samples for direct comparison.

| **Dataset** | **Venue** | **Size** | **DeepFake Types** | **Data Modality** | | | **Answer** | | | **Real Pair** | **Difficulty Level** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | **Img** | **Vid** | **Aud** | **Au** | **Ex** | **Lo** | | |
| FFHQ [48] [link] | [CVPR'19] | 70k | 1 | ✓ | - | - | ✓ | - | - | ✗ | ✗ |
| FakeSpotter [96] [link] | [IJCAI'20] | 12k | 4 | ✓ | - | - | ✓ | - | - | ✗ | ✗ |
| CNNSpot [98] [link] | [CVPR'20] | 724k | 1 | ✓ | - | - | ✓ | - | - | ✗ | ✗ |
| ASVspoof 2019 [101] [link] | [CSL'20] | 108k | 2 | - | - | ✓ | ✓ | - | - | ✓ | ✗ |
| ForgeryNet [34] [link] | [CVPR'21] | 3.1M | 4 | ✓ | ✓ | - | ✓ | - | - | ✓ | ✗ |
| UniversalFake [74] [link] | [CVPR'23] | 720k | 1 | ✓ | - | - | ✓ | - | - | ✗ | ✗ |
| Fake2M [67] [link] | [NeurIPS'23] | 2.3M | 1 | ✓ | - | - | ✓ | - | - | ✗ | ✗ |
| Genimage [141] [link] | [NeurIPS'23] | 2.6M | 1 | ✓ | - | - | ✓ | - | - | ✗ | ✗ |
| Chameleon [116] [link] | [ICLR'25] | 11k | 1 | ✓ | - | - | ✓ | - | - | ✗ | ✗ |
| DeepfakeBench [118] [link] | [NeurIPS'23] | 2M | 1 | - | ✓ | - | ✓ | - | - | ✗ | ✗ |
| GenVideo [8] [link] | [arXiv'24] | 211k | 1 | - | ✓ | - | ✓ | - | - | ✓ | ✗ |
| FakeMusicCaps [16] [link] | [arXiv'24] | 27k | 1 | - | - | ✓ | ✓ | - | - | ✓ | ✗ |
| FakeBench [56] [link] | [arXiv'24] | 3.6k | 1 | ✓ | - | - | ✓ | ✓ | - | ✗ | ✗ |
| FakeClue [105] [link] | [arXiv'25] | 100k | 1 | ✓ | - | - | ✓ | ✓ | - | ✗ | ✗ |
| IVY-FAKE [133] [link] | [arXiv'25] | 160k | 1 | ✓ | ✓ | - | ✓ | ✓ | - | ✗ | ✗ |
| MMFakeBench [65] [link] | [arXiv'25] | 11k | 1 | ✓ | - | - | ✓ | ✓ | - | ✗ | ✗ |
| VIPBench [58] [link] | [arXiv'25] | 80k | 2 | ✓ | - | - | ✓ | ✓ | - | ✓ | ✗ |
| SynthScars [47] [link] | [arXiv'25] | 12k | 1 | ✓ | - | - | ✓ | ✓ | ✓ | ✗ | ✗ |
| MMTD-Set [115] [link] | [ICLR'25] | 153k | 3 | ✓ | - | - | ✓ | ✓ | ✓ | ✗ | ✗ |
| SID-Set [40] [link] | [CVPR'25] | 300k | 2 | ✓ | - | - | ✓ | ✓ | ✓ | ✗ | ✗ |
| LOKI [124] [link] | [ICLR'25] | 18k | 3 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✓ |
| MMIDBench (Ours) [link] | - | 10k | 6 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 3.2 summarizes key public datasets, highlighting their modalities, evaluation tasks, and forgery methods. To further support explainability, several datasets, such as FakeBench, FakeClue, and LOKI, include ground-truth labels for explainability (Ex) and manipulation localization (Lo), enabling researchers to better understand the nature and location of manipulations. However, most datasets lack comprehensive coverage across all data modalities (image, video, and audio) and evaluation tasks, limiting their utility for multi-modal DeepFake detection. Notably, LOKI stands out as one of the few datasets that provides annotations for all three ground-truth labels (authenticity, explainability, and localization) across multiple modalities, making it a valuable resource for holistic evalua-


doi:10.6342/NTU202503386

tion. Despite these advancements, there remains a need for datasets that integrate diverse forgery types, paired real and fake samples, and multi-modal data to comprehensively address the challenges of DeepFake detection and analysis.

The proliferation of expert models and specialized methods for DeepFake generation has significantly heightened the severity of real-world threats associated with synthetic media. Unlike generic, fully synthetic approaches that often produce detectable artifacts or lower-quality forgeries, expert-driven techniques leverage advanced manipulation strategies, such as high-fidelity face swapping, voice cloning, and nuanced behavioral mimicry, to curate highly convincing and targeted DeepFakes. These sophisticated forgeries can evade conventional detection systems and are increasingly being used in high-stakes scenarios, including political disinformation campaigns, financial fraud, and the fabrication of evidence in legal contexts. The ability of expert models to tailor DeepFakes to specific individuals or situations amplifies the potential for reputational damage, social engineering attacks, and large-scale misinformation.

Moreover, as these expert methods become more accessible and user-friendly, the barrier to entry for malicious tools is dramatically reduced. The resulting escalation in both the quality and quantity of DeepFakes poses a profound challenge to digital trust, personal privacy, and the integrity of information ecosystems. As a result, the real-world impact of DeepFakes generated by expert models is both immediate and far-reaching, necessitating urgent advancements in detection, verification, and public awareness.

However, current DeepFake detection benchmarks reveal two major obstacles to building robust and generalizable detectors. The first is the limited diversity of forgery techniques in existing datasets. Many widely used benchmarks, such as FaceForensics+


doi:10.6342/NTU202503386

+ [82], are based on a small set of outdated generation methods. Consequently, models trained on these datasets often fail to detect the more sophisticated manipulations seen in real-world scenarios, leading to a significant drop in performance and highlighting the gap between academic benchmarks and practical effectiveness.

The second challenge is the fragmented approach to evaluation across different modalities and tasks. Most benchmarks focus on a single type of media, such as images, audio, or video, and typically assess only one aspect of explainability, such as narrative reasoning or artifact localization, but rarely both.

To address these issues, we introduce the Multimodal Interpretable Deepfake Benchmark (MMIDBench), a comprehensive new benchmark aimed at advancing DeepFake detection research. As detailed in Table 3.2, MMIDBench introduces several key innovations that directly tackle the field's most urgent challenges.

First, MMIDBench increases forgery diversity by including content generated from 6 distinct DeepFake techniques. These techniques cover a wide range, including Face Swapping, Identity-Preserving Editing, Lip-Syncing, Talking-Head Synthesis, Voice Conversion, Voice Cloning. This broad spectrum of manipulations creates a much more challenging evaluation environment, providing a rigorous test of a detector's ability to generalize to both familiar and novel forgeries. This directly addresses the need for evaluation protocols that keep pace with the rapid evolution of generative AI.

Second, MMIDBench offers a unified, multimodal, and multi-task evaluation framework. It encompasses images, videos, and audio, and integrates three core tasks: authenticity verification, explainability through reasoning or description, and localization of manipulated regions. By combining these elements, MMIDBench enables a comprehensive


doi:10.6342/NTU202503386

assessment of detection models, bridging the gap between generative and discriminative approaches to explainability and encouraging the development of models that perform well across all aspects of detection.

Finally, MMIDBench is a large-scale, balanced dataset, containing 10k samples with a matched “Real Pair” for every fake instance. By addressing both the lack of forgery diversity and the fragmentation of evaluation, it emphasizes that future state-of-the-art models must be generalists, capable of handling a wide variety of forgery types and providing clear, trustworthy explanations across different media. MMIDBench is designed to drive research toward the development of robust, generalizable, and interpretable detection systems.

## 3.2 Real Data and Deepfake Synthesis

The primary goal of this research is to systematically develop a large-scale, multimodal, and diverse benchmark of synthetic media artifacts. This benchmark is not an end in itself but is meticulously constructed to facilitate a rigorous and fair evaluation of contemporary and future DeepFake detection and reasoning methods. The methodology is guided by three core principles that ensure the benchmark remains relevant, diverse, and scientifically robust.

First, the synthesis process exclusively uses a curated set of state-of-the-art generative models. Only models released within the last five years are included, ensuring the generated artifacts reflect the latest technological advancements—from early methods [48] based on Generative Adversarial Networks (GANs) to today’s advanced Diffusion Models [46].


doi:10.6342/NTU202503386

Second, the benchmark emphasizes diversity, not only in the generated data but also in the selection of generative models with fundamentally different architectures. The chosen models span a range of technologies, including GANs [29], Denoising Diffusion Probabilistic Models (DDPMs)[36], and 3D-aware neural rendering frameworks[70]. This architectural variety is crucial, as each model family produces forgeries with unique and recognizable "fingerprints" or artifacts. By incorporating this diversity, the benchmark provides a more comprehensive and challenging evaluation environment, testing detection models on their ability to generalize across different generative approaches. This strategy goes beyond visual realism, aiming to capture the nuanced differences in forgery creation and better reflect the rapidly evolving landscape of synthetic media.

Third, and most importantly, the entire synthesis process is grounded in strict data hygiene protocols to prevent information leakage. Data leakage, where a model is tested on data it has already encountered during training, can seriously undermine the validity of benchmark results. By rigorously preventing this, the protocol ensures that all evaluations are fair, unbiased, and accurately measure how well detection models perform on unseen data.

Given the rapid pace of generative model development, static benchmarks can quickly become outdated. Therefore, this benchmark is built on a principled foundation that anticipates future evaluation needs. By combining modern, architecturally diverse models with strict data hygiene, the benchmark remains both current and methodologically sound. This sets a higher standard for the field, ensuring that progress in DeepFake detection is measured against a challenging and fair benchmark. The breadth and diversity of the benchmark are further detailed in Table 3.3, which summarize the selected DeepFake modalities, manipulation tasks, synthesis methods, and public datasets used.


doi:10.6342/NTU202503386

Table 3.3: Overview of DeepFake modalities (image and video), manipulation tasks, representative synthesis methods, and commonly used public datasets. For each task, we select three publicly available code repositories to ensure diversity in generation approaches. To maintain fair evaluation and avoid data leakage, only the testing splits of public datasets or datasets not used for training are employed for generation.

| Modality | Tasks | Synthesis Method | Public Dataset |
|---|---|---|---|
| Image | Face Swapping | DiffSwap [138] [link] | FaceForensics++ [82]<br>CelebAMaskHQ [52]<br>FFHQ [48] |
| | | CSCS [41] [link] | |
| | | BlendFace [84] [link] | |
| | Identity-Preserving Editing | Step1X-Edit [63] [link] | Emu Edit [83]<br>GEdit-Bench [63]<br>ImgEdit [125] |
| | | MIGE [92] [link] | |
| | | OmniGen2 [107] [link] | |
| Video | Lip-Syncing | DINet [136] [link] | LRS2 [87]<br>VoxCeleb2 [14]<br>TalkingHead-1KH [100] |
| | | LatentSync [53] [link] | |
| | | MuseTalk [135] [link] | |
| | Talking-Head Synthesis | SadTalker [132] [link] | TalkingHead-1KH [100]<br>CelebV-Text [127]<br>HDTF [137] |
| | | Hallo2 [20] [link] | |
| | | AniPortrait [104] [link] | |
| Audio | Voice Cloning | XTTS [1] [link] | EMIME [106]<br>LibriTTS [128]<br>VCTK [95] |
| | | OpenVoice [79] [link] | |
| | | CosyVoice 2.0 [26] [link] | |
| | Voice Conversion | Diff-HierVC [12] [link] | LibriSpeech [76]<br>LibriTTS [128]<br>VCTK [95] |
| | | SpeechT5_VC [3] [link] | |
| | | Seed-VC [62] [link] | |

**Image-Based Forgery Generation.** As demonstrated in Table 3.3, the creation of realistic image-based forgeries focuses on two main manipulation tasks: identity swapping and precise identity-preserving editing. To ensure diversity and challenge, a range of state-of-the-art models, each representing a different architectural approach, are applied to high-quality, publicly available datasets.

- **Face Swapping** involves transferring the identity from a source image onto a target image while preserving the target's attributes like pose, expression, and lighting. DiffSwap [138] reformulates the task as a conditional inpainting problem within a powerful Diffusion Model framework, which excels at creating high-fidelity re-


doi:10.6342/NTU202503386

sults and preserving background details. CSCS [41] introduces a novel training paradigm that generates pseudo-paired data to provide explicit supervision, significantly improving identity preservation where other methods might fail. Finally, BlendFace [84] addresses the subtle but critical issue of identity-attribute entanglement, where models incorrectly associate features like hairstyles with identity, by pre-training its identity encoder on synthetically blended images to learn a more disentangled representation.

- **Identity-Preserving Editing** emphasizes human-centric modifications, specifically altering all elements while preserving the individual's identity. Step1X-Edit [63] focuses on releasing a top open-source image editing model to rival closed-source ones by employing MLLM for processing images and instructions, integrating them with a diffusion decoder, and training on a custom high-quality dataset. MIGE [92] focuses on a unified framework for subject-driven generation and instruction-based editing through standardizing tasks with multimodal instructions and joint training, which enhances consistency, generalization, and cross-task knowledge transfer while excelling in novel compositional tasks. OmniGen2 [107] focuses on a versatile open-source generative model for diverse tasks like text-to-image, editing, and in-context generation via distinct decoding pathways, a decoupled tokenizer, and reflection mechanisms.

**Video-Based Forgery Generation**. As shown in Table 3.3, video manipulation introduces the challenge of temporal consistency, requiring forgeries to be convincing across frames and over time. The benchmark covers two video forgery tasks, each using state-of-the-art models and high-quality public datasets. These tasks span a range of manipulations, including lip-syncing, and talking-head synthesis, reflecting the diversity of real-world


doi:10.6342/NTU202503386

DeepFake scenarios.

- **Lip-Syncing** is an audio-visual task that modifies a video's lip movements to match a new, unrelated audio track. DINet [136] performs this by spatially deforming features from reference images, which helps preserve high-frequency facial textures and avoid the blurriness common in direct-generation methods. LatentSync [53] is an end-to-end latent Diffusion Model that directly models the complex correlations between audio and visual data, introducing a novel temporal alignment mechanism to enhance consistency and reduce flickering artifacts. MuseTalk [135] is a real-time model trained in a Variational AutoEncoder's latent space, designed for efficiency and integration into complete generation pipelines.

- **Talking-Head Synthesis** focuses on generating a complete, animated talking head video from a single source image, typically driven by an audio signal. This is a challenging one-shot task that requires the model to synthesize realistic head motion, facial expressions, and accurate lip movements simultaneously. SadTalker [132] generates talking heads by first learning decoupled 3D motion coefficients for expression and head pose from audio, and then using these coefficients to implicitly modulate a novel 3D-aware face renderer. Hallo2 [20] extends latent diffusion-based models to generate hour-long, 4K resolution animations by introducing enhancements like patch-drop augmentation and vector quantization of latent codes. AniPortrait [104] employs a two-stage framework that first extracts 3D facial mesh and head pose information from audio to create a 2D landmark sequence, which then drives a robust Diffusion Model to synthesize the final photorealistic animation.

**Audio-Based Forgery Generation**. As illustrated in Table 3.3, audio manipulation


doi:10.6342/NTU202503386

presents unique challenges, such as preserving speaker identity, prosody, and naturalness while generating convincing forgeries. The benchmark encompasses two primary audio forgery tasks: voice cloning and voice conversion. For each task, we select three representative, publicly available synthesis methods to ensure diversity in generation approaches.

- **Voice Cloning** is an advanced AI technique that replicates a speaker's unique voice characteristics from minimal audio input. OpenVoice [79] stands out for its versatile, zero-shot cross-lingual cloning that allows granular control over emotions, accents, and intonations without needing extensive training data for new languages. CosyVoice 2.0 [26] leverages discrete speech tokens and optimized models like finite-scalar quantization and chunk-aware flow matching to deliver high-naturalness streaming synthesis with minimal latency, supporting multilingual in-context learning on large datasets. XTTS [1], a deep learning-based TTS model from the Coqui toolkit, excels in zero-shot voice cloning from brief audio clips across 17 languages.

- **Voice Conversion** is a speech processing technique altering the timbre, pitch, or style of a source speaker's voice to match that of a target speaker while preserving the original linguistic content. Diff-HierVC [12] introduces a hierarchical system based on two diffusion models, enhanced by a source-filter encoder and masked priors to boost voice style transfer and speaker adaptation in zero-shot scenarios. Seed-VC [62] tackles zero-shot challenges through an external timbre shifter to prevent leakage and a diffusion transformer for fine-grained timbre capture via in-context learning, achieving superior speaker similarity and low word error rates. SpeechT5_VC [3] proposes a unified encoder-decoder framework with modal-specific pre/post-nets and cross-modal vector quantization for self-supervised speech/text representation learning, demonstrating strong performance.


doi:10.6342/NTU202503386

## 3.3 Multimodal Artifacts Annotation

The rapid advancement of AI has revolutionized digital media, enabling the creation of synthetic content, such as DeepFakes, with unprecedented realism and complexity. As generative techniques have advanced from early GANs to sophisticated Diffusion Models, their artifacts have become increasingly subtle and diverse, posing new challenges for detection systems across images, video, and audio.

This progress has underscored the need for credibility and transparency in detection. In high-stakes domains like law and journalism, simply labeling content as real or fake is insufficient. Modern detection models must provide clear, explainable reasoning, pinpointing specific anomalies and grounding their conclusions in principles that experts can scrutinize. As a result, the field is shifting away from opaque, black-box classifiers toward interpretable frameworks that emphasize transparency and human-like reasoning. To support this shift, the research community has introduced benchmarks such as DD-VQA [134], SID-Set [40], FakeClue [105], MMTD-Set [115], FakeBench [56], and LOKI [124], all designed to promote interpretability and rigorous evaluation.

Notably, LOKI addresses the scalability limitations of earlier benchmarks through a semi-automated, human-in-the-loop annotation process. Leveraging advanced MLLMs such as GPT-4o, LOKI first generates preliminary annotations, identifying anomalous regions and drafting explanations, which are then carefully reviewed and refined by human experts. This hybrid approach combines the efficiency of machine-generated labels with the precision of human oversight, enabling the creation of large-scale, richly annotated datasets and marking a significant methodological advancement in the field.


doi:10.6342/NTU202503386

Despite these advances, a significant gap remains: there is still no unified, systematic approach for annotating artifacts that distinctly evaluates the cognitive abilities of perception and reasoning in detection models. Furthermore, current evaluation practices often use explanations generated by models like GPT-4o as ground truth for assessing all state-of-the-art MLLMs, including GPT-4o itself. This introduces bias, as models tend to favor their own outputs [7].

To address these challenges, we propose a new annotation pipeline and taxonomy that offer a more granular, diagnostic, and scalable framework for evaluating DeepFake detection systems, while also reducing model bias in assessment. Drawing inspiration from the question-driven detail of DD-VQA, our approach introduces a structured taxonomy that categorizes artifacts by their scope and the cognitive domain they test. By distinguishing whether artifact detection relies on low-level perception or high-level reasoning, our framework is capable of supporting separate diagnostic labels for each capability, enabling a comprehensive cognitive profile of detection models. This detailed evaluation is essential for guiding targeted research and accelerating the development of DeepFake detectors that are both accurate and robustly intelligent.

For each DeepFake generation technique, we generate at least 100 samples to form one-to-one real-fake pairs, resulting in a total of 600 pairs spanning three different modalities. To fully evaluate both perceptual and reasoning capabilities, we annotate these samples using our proposed taxonomy, which distinguishes between quality artifacts, issues affecting overall media integrity, and semantic artifacts, errors with meaningful implications that require precise localization. This comprehensive annotation not only enables a nuanced assessment of detection models but also facilitates a deeper understanding of the strengths and limitations of current generative techniques across diverse scenarios.


doi:10.6342/NTU202503386

**Quality Artifacts** refer to issues that affect the overall integrity of the media, such as blur, compression artifacts, or flickering. These artifacts may occur across the entire media or in manipulated areas. However, their detection does not rely on identifying a specific semantic location, as their impact is primarily on the perceptual quality of the content.

**Semantic Artifacts**, in contrast, are errors that have meaningful implications and require identifying their specific location within the media. These artifacts often result from inconsistencies in the generative process or violations of real-world principles. Identifying their location enhances interpretability, making the reasoning behind DeepFake detection more transparent and understandable.

As summarized in Table 3.4, our taxonomy classifies artifacts based on their characteristics and the cognitive skills required for detection, distinguishing between general quality-related artifacts and those with semantic significance that require precise localization. By categorizing artifacts by type, providing concise descriptions, and linking each tag to specific cognitive skills, this structured approach supports detailed annotation and forensic analysis, offering deeper insights into the strengths and limitations of generative models.

**Annotation Pipeline in Practice**. To implement the unified taxonomy at scale, we have developed a dedicated annotation platform optimized for hierarchical classification and robust quality assurance. The annotation process is fully manual and human-driven, prioritizing accuracy and reliability over automation. In light of the current limitations of MLLMs, such as the 59% accuracy ceiling observed with GPT-4o on DeepFake detection tasks, as reported by LOKI, we have intentionally excluded AI-assisted pre-annotation.


doi:10.6342/NTU202503386

All annotations are produced directly by trained human annotators, who follow detailed guidelines and possess a deep understanding of the unified taxonomy.

A key feature of our platform is the side-by-side comparison of real-fake media pairs, each matched in a strict one-to-one correspondence. This layout enables annotators to systematically compare manipulated samples with their authentic counterparts, facilitating the precise identification of both quality and semantic artifacts. The interface supports this comparative analysis with synchronized playback, zoom, and frame-by-frame navigation, while structured checklists and dropdown menus allow annotators to assign taxonomy-based labels at multiple levels of granularity.

To maintain the highest standards of annotation quality, each annotated instance is independently reviewed by at least two annotators. Any discrepancies or disagreements are automatically flagged for adjudication by a senior annotator, ensuring consensus and consistency. Additionally, a rigorous quality control mechanism is embedded throughout the workflow: curated "golden" questions, samples with known ground-truth answers, are randomly injected into the annotation stream. Annotator performance on these questions is continuously monitored, and if accuracy falls below a predefined threshold, the relevant annotations are manually reviewed and, if necessary, reannotated. This comprehensive, human-centered approach ensures the creation of high-quality, gold-standard datasets.


doi:10.6342/NTU202503386

Table 3.4: Taxonomy of General Artifacts in Media Content.

| **Category** | **Tags** | **Descriptions** | **Location** |
|---|---|---|---|
| Quality Artifacts | Blur or Compression Artifacts | The sample appears out of focus or lacks sharpness, or shows visible distortions or noise patterns from digital compression, such as color banding or pixelation. The overall visual impression is soft, unclear, or degraded. | |

Continued on next page



doi:10.6342/NTU202503386

Table 3.4 (continued): Taxonomy of General Artifacts in Media Content

| Category | Tags | Descriptions | Location |
|---|---|---|---|
| Quality Artifacts | Unnatural Texture | The skin appears excessively smooth or lacks natural fine details. Features such as hair, eyebrows, or teeth may look blurry, artificial, or unrecognizable. The image may also contain unrecognizable words or abnormal backgrounds that do not match the context. | |

Continued on next page



doi:10.6342/NTU202503386

Table 3.4 (continued): **Taxonomy of General Artifacts in Media Content**

| **Category** | **Tags** | **Descriptions** | **Location** |
|---|---|---|---|
| Quality Artifacts | Motion or Temporal Inconsistency | The movement or appearance of objects, subjects, or the background across consecutive frames appears unnatural, lacks continuity, or shows rapid, unstable fluctuations in texture, color, or geometry. This includes mismatched frame transitions, unrealistic motion, or flickering effects that cause an unstable or shimmering appearance. | |

Continued on next page



doi:10.6342/NTU202503386

**Table 3.4 (continued): Taxonomy of General Artifacts in Media Content**

| **Category** | **Tags** | **Descriptions** | **Location** |
| --- | --- | --- | --- |
| Quality Artifacts | Audio Glitches | Unnatural sounds in the audio, including clicks, pops, metallic noises, or digital artifacts that disrupt the listening experience. | |
| Semantic Artifacts | Inconsistent Lighting | The brightness or lighting of the subject does not match the environment, making the subject appear unnaturally lit or out of place. | |
| Semantic Artifacts | Unnatural Expression | The facial expression appears unnatural, rigid, exaggerated, or contextually inappropriate. | |
| Semantic Artifacts | Inconsistent Reflections of Glasses | Reflections on eyeglasses are absent, distorted, or mismatched, violating physical principles. | |

Continued on next page



doi:10.6342/NTU202503386

**Table 3.4 (continued): Taxonomy of General Artifacts in Media Content**

| Category | Tags | Descriptions | Location |
|---|---|---|---|
| Semantic Artifacts | Anatomical Incoherence | Unnatural proportions, misalignment, or movements of body parts or facial features. | Eyebrows, Eyes, Nose, Mouth, Cheek, Chin, Teeth, Tongue, Neck, Shoulders, Arms, Hands, Chest, Waist, Hips, Thighs, Knees, Feet |
| Semantic Artifacts | Age Misalignment | There is a noticeable mismatch between the apparent ages of different features in the image. | Hair, Eyebrows, Eyes, Nose, Mouth, Cheek, Chin |
| Semantic Artifacts | Visible Boundary or Edge | A visible seam, blur, or color mismatch appears at the boundary where the altered facial region joins the rest of the image. | Hairline, Forehead, Cheek, Chin |

Continued on next page



doi:10.6342/NTU202503386

Table 3.4 (continued): Taxonomy of General Artifacts in Media Content

| Category | Tags | Descriptions | Location |
|---|---|---|---|
| Semantic Artifacts | Accessory Artifacts | Glasses, earrings, hats, or other accessories may be rendered inconsistently, disappear, or become distorted during manipulation. | Hair, Eyebrows, Eyes, Ears, Nose |
| Semantic Artifacts | Unnatural Prosody | Speech often sound robotic, monotonous, or flat, lacking natural intonation. | |
| Semantic Artifacts | Mispronunciation | Mumbled, unclear, or nonsensical sounds that will make it unrecognizable. | |

## 3.4 Assertion-Based Evaluation and Question Generation

To rigorously assess the DeepFake detection capabilities of MLLMs, we introduce a novel assertion-based evaluation framework. This framework combines several question types, each targeting different aspects of model perception and reasoning:

**Judgment Tasks**. MLLMs are asked to determine whether a given sample is real or a DeepFake. To minimize prompt bias, we pose the question in two forms: (a) asking if


doi:10.6342/NTU202503386

the sample is AI-synthesized or real, and (b) asking the model to identify which sample is real or AI-generated. We further categorize task difficulty based on human performance: *easy* if all human evaluators answer correctly, *hard* if more than half answer incorrectly, and *medium* otherwise.

**Multiple Choice Tasks**. MLLMs are presented with paired real and synthetic data from the same domain and must select which sample is a DeepFake. This format leverages our curated real-synthetic pairs to test the model's comparative analysis skills.

**DeepFake Reasoning**. We include open-ended questions that require MLLMs to explain why a sample is identified as DeepFake-generated. This assesses the model's ability to ground its reasoning in observable details, rather than relying on guesswork.

**Tag-Based Assertions**. Since most MLLMs lack the ability to provide pixel-level predictions, we instead utilize semantic-level tags to evaluate their perceptual capabilities. We begin by defining a set of mutually exclusive tags, such as specific regions or the entire image, video, or audio, which are annotated in advance for each sample. Using these predefined tags, we systematically generate prompts that require models to identify and localize artifacts, covering a spectrum from low-level quality defects to high-level semantic inconsistencies. For each perspective of perception ability (e.g., detection of compression artifacts, unnatural textures, or semantic anomalies), GPT-4o is prompted to generate targeted assertions based on the annotated tags. All generated assertions undergo manual review to ensure both accuracy and clarity. This approach enables a comprehensive and fine-grained assessment of MLLMs' abilities to perceive and reason about artifacts, while also streamlining the annotation process.


doi:10.6342/NTU202503386

# Chapter 4 Experiments

In this section, we assess a range of MLLMs using our newly developed MMIDBench evaluation framework. This assessment covers both open-source and commercial models, with all tests performed in a zero-shot manner. The subsections that follow begin by describing the models we evaluated and the procedures we followed. We then examine how well current MLLMs perform on DeepFake detection, drawing comparisons to human participants and specialized expert models. Finally, we explore the key difficulties and limitations that MLLMs encounter today.

## 4.1 Baselines

**MLLMs**. We assessed a total of 7 proprietary models, which vary in their architectures, sizes, and capabilities, spanning vision, language, audio, and video modalities. The proprietary models encompass GPT-4.1, GPT-4o, GPT-4.1 mini, GPT-4o mini, GPT-4.1 nano [42], Gemini-2.5-Pro ,and Gemini-2.5-Flash [17]. All of these support inputs in vision and language, and Gemini-2.5-Pro additionally handles audio.

**Human Evaluation**. We recruited more than 10 participants, consisting of advanced university students and senior graduates, to take part in evaluation tasks. These tasks involved judgment assessments and multiple-choice quizzes focused on various types of


doi:10.6342/NTU202503386

DeepFake data. To promote reliable outcomes, we had at least three participants review each item.

## 4.2 Assertion-Based Evaluation

In this section, we formalize the assertion-based grading logic used to evaluate model responses against ground truth artifacts in multimedia deepfake detection tasks. The evaluation quantifies how effectively a model identifies and verifies predefined perceptual artifacts (e.g., unnatural textures, inconsistent lighting) in its natural language analysis. We define two key metrics: the Sample Perception Verification Score (SPVS) for individual samples and the Overall Perception Verification Score (OPVS) for aggregate performance. These metrics rely on keyword matching to detect verified assertions in model responses, with special handling for edge cases such as refusals. Note that the benchmark exclusively uses fake samples for open-ended questions, ensuring that positive assertions (artifacts) are always present; thus, cases where no artifacts exist are not applicable.

Let $\mathcal{T}$ denote the set of tasks (e.g., image editing, audio voice cloning). For each task $t \in \mathcal{T}$, let $\mathcal{S}_t$ be the set of samples, where each sample $s \in \mathcal{S}_t$ consists of a file path, a model response $r_s$ (a natural language string), and a ground truth vector $\mathbf{g}_s = (g_{s,1}, \ldots, g_{s,P}) \in \{0,1\}^P$. Here, $P$ is the number of perception perspectives (artifact categories), and $g_{s,p} = 1$ if perspective $p$ has a positive assertion (i.e., the artifact exists) for sample $s$, else $0$.

The keyword mapping is defined as a function $K : \{1, \ldots, P\} \to 2^{\Sigma}$, where $\Sigma$ is the vocabulary of strings, and $K(p)$ yields a set of case-insensitive keywords associated with perspective $p$ (e.g., for "Unnatural Texture", $K(p) = \{$“texture”, “waxy”, $\ldots\}$).



doi:10.6342/NTU202503386

An assertion for perspective $p$ is verified in response $r_s$ if there exists a keyword $k \in K(p)$ such that $k$ appears in the lowercase version of $r_s$. Define the verification indicator as:

$$v_{s,p} = \begin{cases} 1 & \text{if } \exists k \in K(p) \text{ s.t. } k \in r_s^{\downarrow}, \\ 0 & \text{otherwise,} \end{cases}$$

where $r_s^{\downarrow}$ denotes the lowercase transformation of $r_s$.

The total positive assertions for sample $s$ is $A_s = \sum_{p=1}^{P} g_{s,p}$, and the verified assertions is $V_s = \sum_{p=1}^{P} v_{s,p} \cdot g_{s,p}$ (counting only matches where $g_{s,p} = 1$). Since only fake samples are evaluated, $A_s > 0$ always holds.

The SPVS for sample $s$ is then computed as:

$$\text{SPVS}_s = \begin{cases} 0 & \text{if } r_s^{\downarrow} \text{ contains a refusal phrase (e.g., "I cannot determine")}, \\ \frac{V_s}{A_s} & \text{otherwise.} \end{cases}$$

For each task $t$, the task-average SPVS is $\overline{\text{SPVS}}_t = \frac{1}{|\mathcal{S}_t|} \sum_{s \in \mathcal{S}_t} \text{SPVS}_s$, provided $|\mathcal{S}_t| > 0$.

Finally, the OPVS aggregates across all tasks and samples:

$$\text{OPVS} = \frac{1}{N} \sum_{t \in \mathcal{T}} \sum_{s \in \mathcal{S}_t} \text{SPVS}_s,$$

where $N = \sum_{t \in \mathcal{T}} |\mathcal{S}_t|$ is the total number of graded samples. This metric ranges from 0 to 1, with higher values indicating better alignment between model perceptions and ground truth artifacts. Refusals penalize the score to encourage decisive analyses.

To ensure the quality and reliability of the generated assertions during the annotation


doi:10.6342/NTU202503386

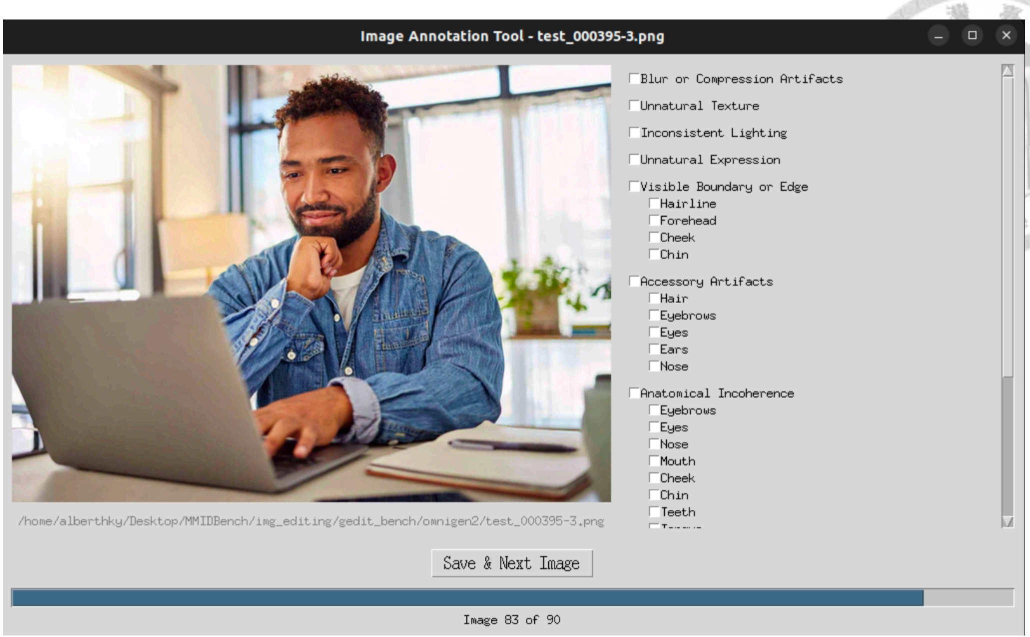


Figure 4.1: **Labeling interface** used for quality assessment. A user examines the sample on the left and uses the checklist on the right to flag any visual artifacts or logical inconsistencies.

process, we developed a specialized labeling interface, as illustrated in Figure 4.1. This interface facilitates manual review by presenting the sample (e.g., an image, video, or audio) on the left panel for close examination, while the right panel features a comprehensive checklist of potential artifacts categorized by type, such as incoherent textures, unnatural lighting, visible boundaries or edges, forehead or chin distortions, and accessory artifacts. Annotators systematically flag any visual or logical inconsistencies within the predefined tags $\mathcal{T}_s$ for each sample $s$, ensuring that assertions in $\mathcal{A}_{s,p}$ are accurate, clear, and free from biases. This human-in-the-loop approach not only validates the outputs from the generator model $g_{\text{Gen}}$ but also enhances the robustness of the perception verification scores by grounding them in expert-verified annotations.

While the current annotation system, as depicted in Figure 4.1, serves as a temporary solution for initial quality assessments and manual reviews, we are actively developing a more advanced web-based tool to support large-scale annotations. This forthcoming platform will enable efficient collaboration among multiple annotators, streamline the review


doi:10.6342/NTU202503386

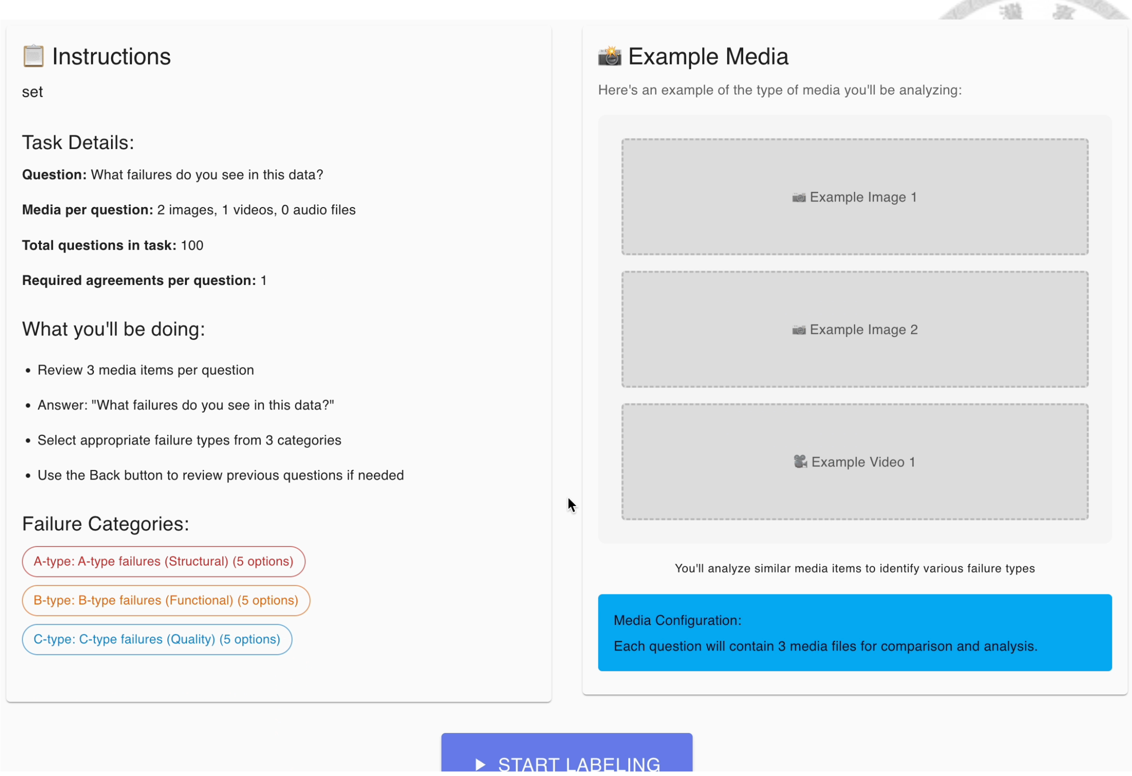


Figure 4.2: Overview of the web-based labeling interface, featuring initial instructions, task details (e.g., total questions across modalities), required agreements, failure type categories, and placeholders for example media to guide annotators before starting the labeling process.

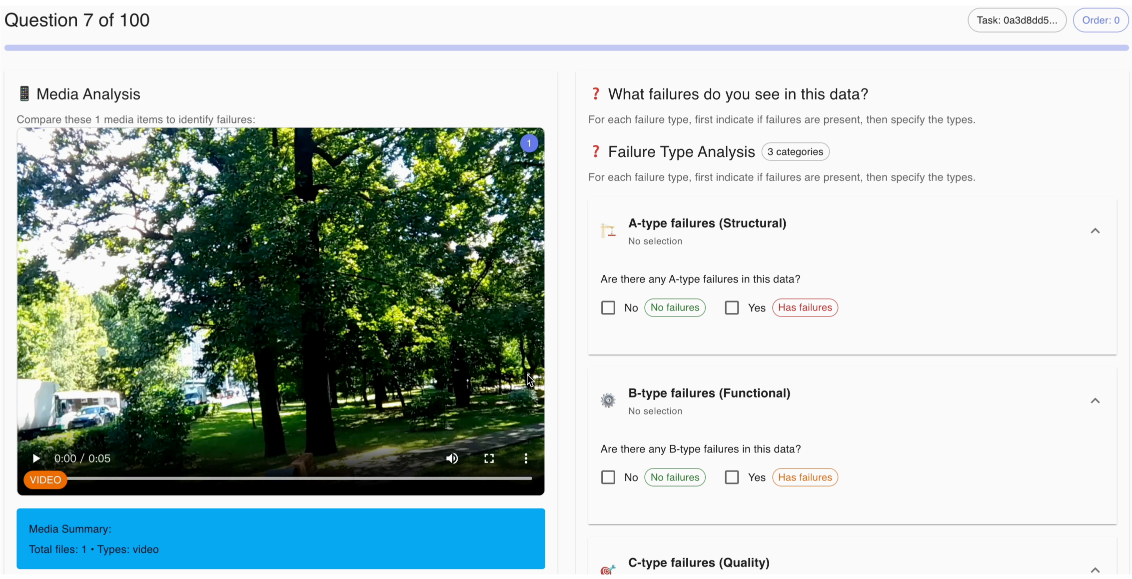


Figure 4.3: Example of the web-based labeling interface in action, displaying a specific annotation task with media playback for analysis, detailed questions on failure types, and interactive Yes/No selection buttons for flagging artifacts or inconsistencies.

and validation processes for assertions across diverse modalities, and incorporate features like dynamic media playback, customizable checklists, and real-time progress tracking to


doi:10.6342/NTU202503386

enhance scalability and accuracy in evaluating MLLM perceptual capabilities, as illustrated in Figures 4.2 and 4.3.

## 4.3 Interpretable DeepFake Detection Results

In this section, we analyze the performance of various MLLMs on the proposed MMIDBench dataset, focusing on their ability to detect and interpret DeepFake content across image, video, and audio modalities. The evaluation encompasses three task types: True-False binary classification, Multiple-Choice questions, and Open-Ended responses. We compare models such as Gemini-2.5-Pro, Gemini-2.5-Flash, GPT-4.1, and GPT-4o against baselines like random choice and human performance. Results reveal significant challenges in DeepFake detection, with models generally underperforming humans, particularly in nuanced tasks requiring reasoning about manipulation techniques.

Table 4.1: Results of different models on the MMIDBench for True-False, Multiple Choice questions, and Open-Ended questions.

| | True-False | | | | Multiple Choice | | | | Open-Ended | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Image | Video | Audio | Overall | Image | Video | Audio | Overall | Image | Video | Audio | Overall |
| Random Choice | 50.2 | 49.9 | 49.6 | 49.9 | 50.3 | 48.7 | 48.3 | 49.1 | - | - | - | - |
| Human | 83.2 | 90.3 | 75.7 | 83.1 | 91.3 | 95.6 | 84.5 | 90.5 | - | - | - | - |
| Gemini-2.5-Pro | 67.3 | 58.6 | **69.2** | 65.0 | 50.7 | 55.0 | **51.7** | **52.5** | **66.1** | **75.6** | **80.3** | **74.0** |
| Gemini-2.5-Flash | 62.9 | **67.5** | 55.8 | 62.1 | **53.5** | **57.5** | 45.8 | 52.1 | 65.2 | 68.0 | 78.4 | 70.5 |
| GPT-4.1 | **76.1** | 67.3 | - | **71.7** | 53.1 | 48.6 | - | 50.9 | 62.6 | 51.9 | - | 57.3 |
| GPT-4o | 57.7 | 49.8 | - | 53.8 | 39.0 | 2.7 | - | 20.9 | 52.8 | 8.5 | - | 30.6 |

**Overall Performance Across Tasks and Modalities**. As illustrated in 4.1, the evaluated models display diverse performance profiles across the True-False, Multiple Choice, and Open-Ended tasks, with the input modalities of image, video, and audio.

In the True-False task, which assesses binary discrimination of manipulated versus authentic media, GPT-4.1 emerges as the top performer with an overall accuracy of 71.7%,


doi:10.6342/NTU202503386

driven by its strong results in image-based detection at 76.1% and a competitive showing in video at 67.3%. Gemini-2.5-Pro ranks second overall at 65.0%, showcasing balanced capabilities, including the highest audio performance at 69.2% and solid image results at 67.3%, though it dips in video to 58.6%. Gemini-2.5-Flash achieves 62.1% overall, with its peak in video at 67.5% but lower scores in audio at 55.8%, suggesting modality-specific strengths possibly tied to its optimized processing for dynamic content. In contrast, GPT-4o trails with 53.8% overall, underperforming in video at 49.8% and showing moderate image accuracy at 57.7%. Notably, none of the models approach human-level performance, which averages 83.1% overall, with humans excelling particularly in video at 90.3% and images at 83.2%; this gap underscores limitations in the models' ability to reliably detect subtle manipulations, even as they surpass random choice baselines around 49.9%. These disparities may stem from differences in training data emphasis, where models like GPT-4.1 benefit from extensive visual pretraining, while others struggle with audio nuances or video temporal dynamics.

Shifting to the Multiple Choice task, which requires selecting the correct manipulation type from predefined options, the models generally hover near or slightly above random chance, revealing challenges in precise categorization. Gemini-2.5-Pro leads with an overall score of 52.5%, bolstered by its audio performance at 51.7% and video at 55.0%, though image results are modest at 50.7%. Gemini-2.5-Flash follows closely at 52.1% overall, achieving the highest scores in images at 53.5% and videos at 57.5%, but lagging in audio at 45.8%, which might indicate weaker integration of auditory signals compared to its visual prowess. GPT-4.1 attains 50.9% overall, with 53.1% in images and 48.6% in videos, barely exceeding the random baseline of 49.1%, while GPT-4o performs abysmally at 20.9% overall, plummeting to 2.7% in videos despite a somewhat better


doi:10.6342/NTU202503386

39.0% in images. The stark underperformance in video for GPT models can be attributed to their lack of native video processing; to mitigate this, we adopted the LOKI methodology by extracting one frame per second, yet the models still failed to effectively leverage these static representations for accurate selection, often defaulting to incorrect or random choices. Human evaluators, by comparison, achieve 90.5% overall, with exceptional accuracy in videos at 95.6% and images at 91.3%, highlighting how models falter in synthesizing multimodal cues under constrained response formats. This task exposes broader limitations in fine-grained reasoning, where even advanced models like Gemini struggle to outperform chance in audio-heavy scenarios, potentially due to insufficient exposure to diverse manipulation artifacts during training.

In the Open-Ended task, demanding detailed explanations of detected manipulations and artifact identification without predefined options, the models demonstrate more pronounced strengths in interpretive capabilities, particularly those with robust multimodal architectures. Gemini-2.5-Pro dominates with an overall score of 74.0%, excelling across all modalities—66.1% in images, 75.6% in videos, and 80.3% in audio—reflecting its advanced integration of temporal and auditory processing that enables richer descriptive outputs. Gemini-2.5-Flash secures second place at 70.5% overall, with strong showings in audio at 78.4% and videos at 68.0%, alongside 65.2% in images, suggesting efficiency in generating coherent explanations for dynamic media. GPT-4.1 reaches 57.3% overall, performing best in images at 62.6% but dropping to 51.9% in videos, with no audio evaluation available, which limits its applicability in fully multimodal contexts. GPT-4o again lags significantly at 30.6% overall, with 52.8% in images but a mere 8.5% in videos, further emphasizing its struggles with non-static inputs despite frame extraction via LOKI. These results illuminate how models like Gemini-2.5-Pro, designed for seam-


doi:10.6342/NTU202503386

less multimodal fusion, better handle the task's demands for nuanced analysis and artifact pinpointing, such as identifying audio splicing or video deepfake inconsistencies. However, the absence of human benchmarks here prevents direct comparison, though the models' outputs often lack the depth of human intuition in articulating subtle cues. Overall, the experiment reveals that while progress in multimodal AI is evident, persistent gaps in video and audio handling—exacerbated by preprocessing needs—point to opportunities for future enhancements in model architectures and training paradigms to bridge the divide with human-level detection and explanation.

**Performance Breakdown by Deepfake Techniques**. Delving deeper into technique-specific results, Tables 4.2 to 4.4 illustrate how models handle distinct DeepFake methods: Face Swapping (FS) and Identity-Preserving Editing (IPE) for images, Lip-Syncing (LS) and Talking-Head Synthesis (THS) for videos, and Voice Cloning (VCl) and Voice Conversion (VCo) for audio.

In the True-False task in Table 4.2, GPT-4.1 excels overall at 71.7%, with top scores in image IPE at 76.7% and video THS at 82.1%. Gemini-2.5-Pro achieves 65.0% overall, leading in audio VCl at 76.7% and VCo at 61.7%, while showing strength in image FS at 74.2%. Gemini-2.5-Flash performs well in video THS at 71.3% and LS at 63.7%, but dips in audio. GPT-4o trails at 53.8% overall, with inconsistent results like 69.6% in image FS but only 41.6% in video LS. These patterns indicate that models are more adept at detecting overt manipulations like FS or THS, where visual inconsistencies are prominent, compared to subtler edits like IPE or VCo.

As illustrated in Table 4.3, the Multiple-Choice evaluation is confined to GPT models and image/video modalities. GPT-4.1 attains an overall score of 50.9%, demonstrating


doi:10.6342/NTU202503386

balanced results across FS at 57.1%, IPE at 49.0%, LS at 50.1%, and THS at 47.1%. In contrast, GPT-4o plummets to 20.8% overall, nearly failing in video-related tasks with scores of 0.6% in LS and 4.81% in THS. This shortfall likely arises from difficulties in handling video frames as substitutes for actual video inputs.

The Open-Ended results in Table 4.4 demonstrate Gemini-2.5-Pro's clear dominance with an overall score of 74.0%, excelling particularly in video THS (79.8%) and audio VCl (81.7%) and VCo (78.9%). Gemini-2.5-Flash follows closely with an overall score of 70.5%, showing strong performance in image IPE (72.2%) and audio VCl (80.6%). GPT-4.1 achieves a respectable 57.3% overall, with its best performance in video THS (57.5%). In contrast, GPT-4o lags significantly with an overall score of 30.6%, struggling particularly in video tasks like LS (7.2%) and THS (9.7%), highlighting its limitations in processing video frames. Across tasks, audio techniques such as VCl and VCo remain challenging due to the nuanced nature of vocal timbre changes, while video THS benefits from models' strengths in motion analysis and temporal consistency.

Table 4.2: Performance Comparison Across DeepFake Techniques, Face Swapping (FS), Id-Preserving Editing (IPE), Lip-Syncing (LS), Talking-Head Synthesis (THS), Voice Cloning (VCl), and Voice Conversion (VCo), on the True-False Task.

| **True-False** | **Image** | | **Video** | | **Audio** | | |
|---|---|---|---|---|---|---|---|
| | **FS** | **IPE** | **LS** | **THS** | **VCl** | **VCo** | **Overall** |
| Gemini-2.5-Pro | 74.2 | 60.4 | 61.3 | 55.8 | **76.7** | **61.7** | 65.0 |
| Gemini-2.5-Flash | 70.0 | 55.8 | **63.7** | 71.3 | 53.3 | 58.3 | 62.0 |
| GPT-4.1 | **75.4** | **76.7** | 52.5 | **82.1** | - | - | **71.7** |
| GPT-4o | 69.6 | 45.8 | 41.6 | 57.9 | - | - | 53.8 |

**Impact of Model Size on Performance**. To assess scalability, Table 4.5 compares model sizes within the GPT family on the True-False task for image and video modalities. Larger models consistently outperform smaller variants; GPT-4.1 achieves 71.7% overall,



doi:10.6342/NTU202503386

Table 4.3: Performance Comparison Across DeepFake Techniques, Face Swapping (FS), Id-Preserving Editing (IPE), Lip-Syncing (LS), Talking-Head Synthesis (THS), Voice Cloning (VCl), and Voice Conversion (VCo), on the Multiple-Choice Task.

| **Multiple-Choice** | **Image** | | **Video** | | **Audio** | | |
|---|---|---|---|---|---|---|---|
| | **FS** | **IPE** | **LS** | **THS** | **VCl** | **VCo** | **Overall** |
| Gemini-2.5-Pro | 53.3 | 48.1 | **56.7** | 53.3 | **53.3** | **50.0** | **52.5** |
| Gemini-2.5-Flash | 53.3 | **52.6** | 50.0 | **65.0** | 48.3 | 43.3 | 52.1 |
| GPT-4.1 | **57.1** | 49.0 | 50.1 | 47.1 | - | - | 50.9 |
| GPT-4o | 37.1 | 40.8 | 0.6 | 4.81 | - | - | 20.8 |

Table 4.4: Performance Comparison Across DeepFake Techniques, Face Swapping (FS), Id-Preserving Editing (IPE), Lip-Syncing (LS), Talking-Head Synthesis (THS), Voice Cloning (VCl), and Voice Conversion (VCo), on the Open-Ended Task .

| **Open-Ended** | **Image** | | **Video** | | **Audio** | | |
|---|---|---|---|---|---|---|---|
| | **FS** | **IPE** | **LS** | **THS** | **VCl** | **VCo** | **Overall** |
| Gemini-2.5-Pro | **62.0** | 70.2 | 71.3 | 79.8 | **81.7** | **78.9** | **74.0** |
| Gemini-2.5-Flash | 58.2 | **72.2** | 64.6 | 71.3 | 80.6 | 76.1 | 70.5 |
| GPT-4.1 | 59.1 | 66.1 | 46.3 | 57.5 | - | - | 57.3 |
| GPT-4o | 43.9 | 61.7 | 7.2 | 9.7 | - | - | 30.6 |

significantly ahead of GPT-4.1 mini (2025) at 54.5%, GPT-4.1 mini (2024) at 46.3%, and GPT-4.1 nano at 45.0%. Similarly, GPT-4o at 53.8% surpasses GPT-4o mini at 46.3%. This trend, evident in image scores dropping from 76.1% (GPT-4.1) to 36.9% (nano) and video from 67.3% to 39.4% (mini versions), suggests that increased parameters enhance multimodal reasoning, though all remain below human levels at 86.8% overall. Smaller models approach random choice (50.1%), implying that computational scale is crucial for detecting DeepFake nuances.

Overall, these experiments demonstrate that while MLLMs show promise in DeepFake detection, particularly in binary and open-ended formats, they lag in multiple-choice scenarios and subtler techniques. Future work could explore fine-tuning for audio modalities and integrating specialized detectors to bridge the human-model gap.


doi:10.6342/NTU202503386

Table 4.5: Performance Comparison Across Model Sizes and Versions on the True-False Task in MMIDBench. This table presents experimental results to analyze how model size influences overall performance.

| **True-False** | **Image** | **Video** | **Overall** |
|---|---|---|---|
| Random Choice | 50.2 | 49.9 | 50.1 |
| Human | 83.2 | 90.3 | 86.8 |
| GPT-4.1 | 76.1 | 67.3 | 71.7 |
| GPT-4.1 mini (2025) | 57.5 | 51.5 | 54.5 |
| GPT-4.1 mini (2024) | 53.1 | 39.4 | 46.3 |
| GPT-4.1 nano | 36.9 | 53.1 | 45.0 |
| GPT-4o | 57.7 | 49.8 | 53.8 |
| GPT-4o mini | 53.1 | 39.4 | 46.3 |

**Qualitative Examples of Deepfake Detection**. For image-based detection, participants are tasked with determining whether a given image is authentic or generated by DeepFake techniques. For instance, in Figure 4.4, an image was identified as a DeepFake due to artifacts such as unnatural, waxy skin textures and poorly rendered hair details. These visual artifacts, which deviate from the sharpness and natural appearance of real photographs, serve as critical evidence for detection. The concise and focused analysis of such artifacts is essential for accurate classification.

In video-based detection, participants analyze two videos to identify which one is most likely generated by DeepFake techniques. A notable example in Figure 4.5 consists of a video where the subject's facial movements appeared unnatural and mask-like. While the mouth moved to form words, the upper parts of the face, such as the cheeks and eyes, remained static and uncoordinated. These inconsistencies in facial dynamics are key indicators of DeepFake generation, emphasizing the importance of analyzing both global and localized motion patterns in videos.

Audio-based detection requires a detailed analysis of auditory cues to determine


doi:10.6342/NTU202503386

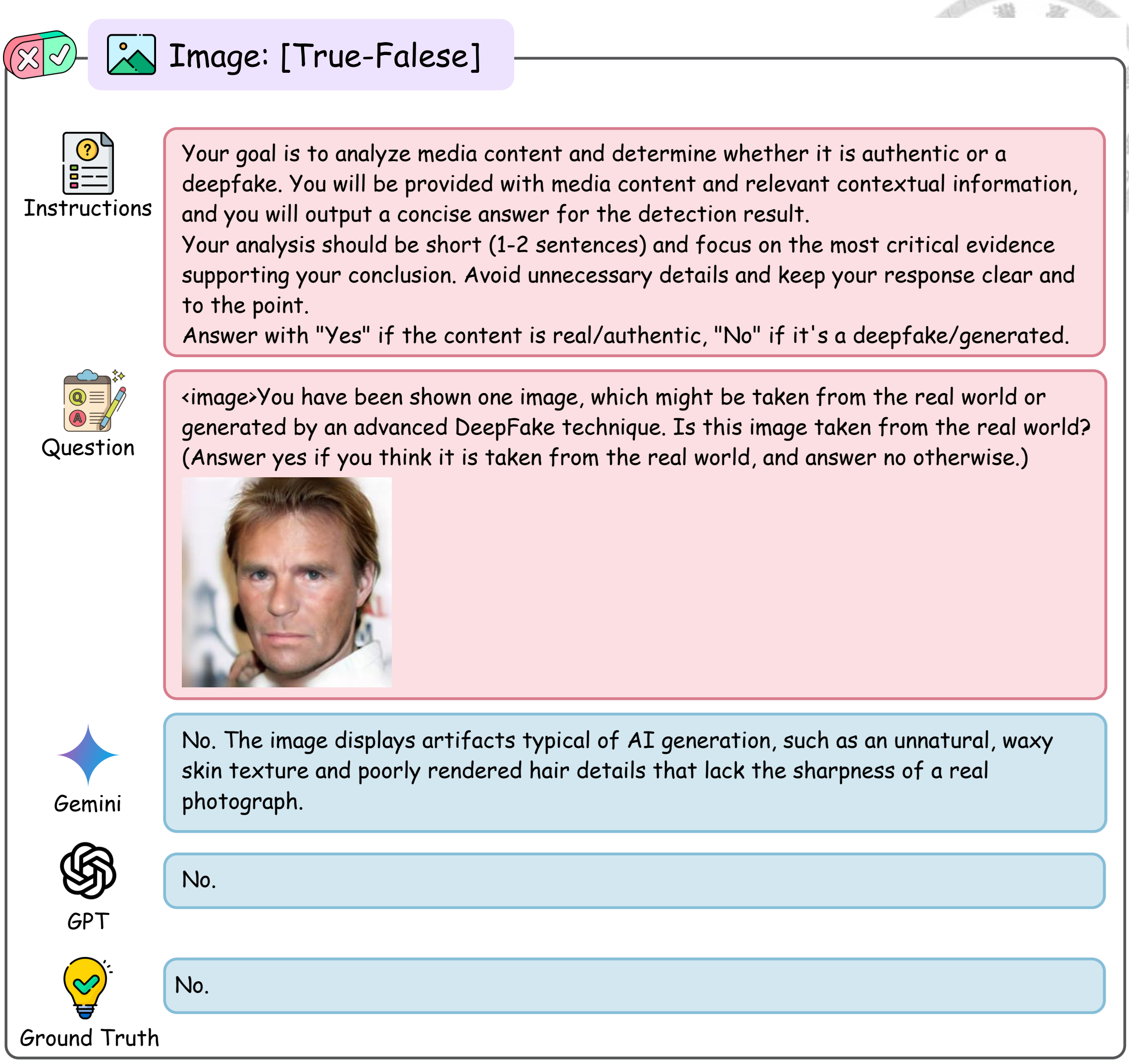


Figure 4.4: A True-False question example in the image modality, showcasing both Gemini's and GPT's responses.

whether a clip is authentic or generated. In Figure 4.6, a DeepFake audio clip was identified based on global auditory cues such as unnatural rhythm, rigid pacing, and flat intonation. The absence of natural human vocalizations, such as breath sounds or slight hesitations, further contributed to the detection. Segmental analysis revealed artificiality in sibilant sounds and vowel prolongations, underscoring the need for a thorough examination of both global and segmental features in audio DeepFake detection.

These examples illustrate the diverse challenges posed by DeepFake detection across different modalities and highlight the critical role of detailed analysis in identifying subtle


doi:10.6342/NTU202503386

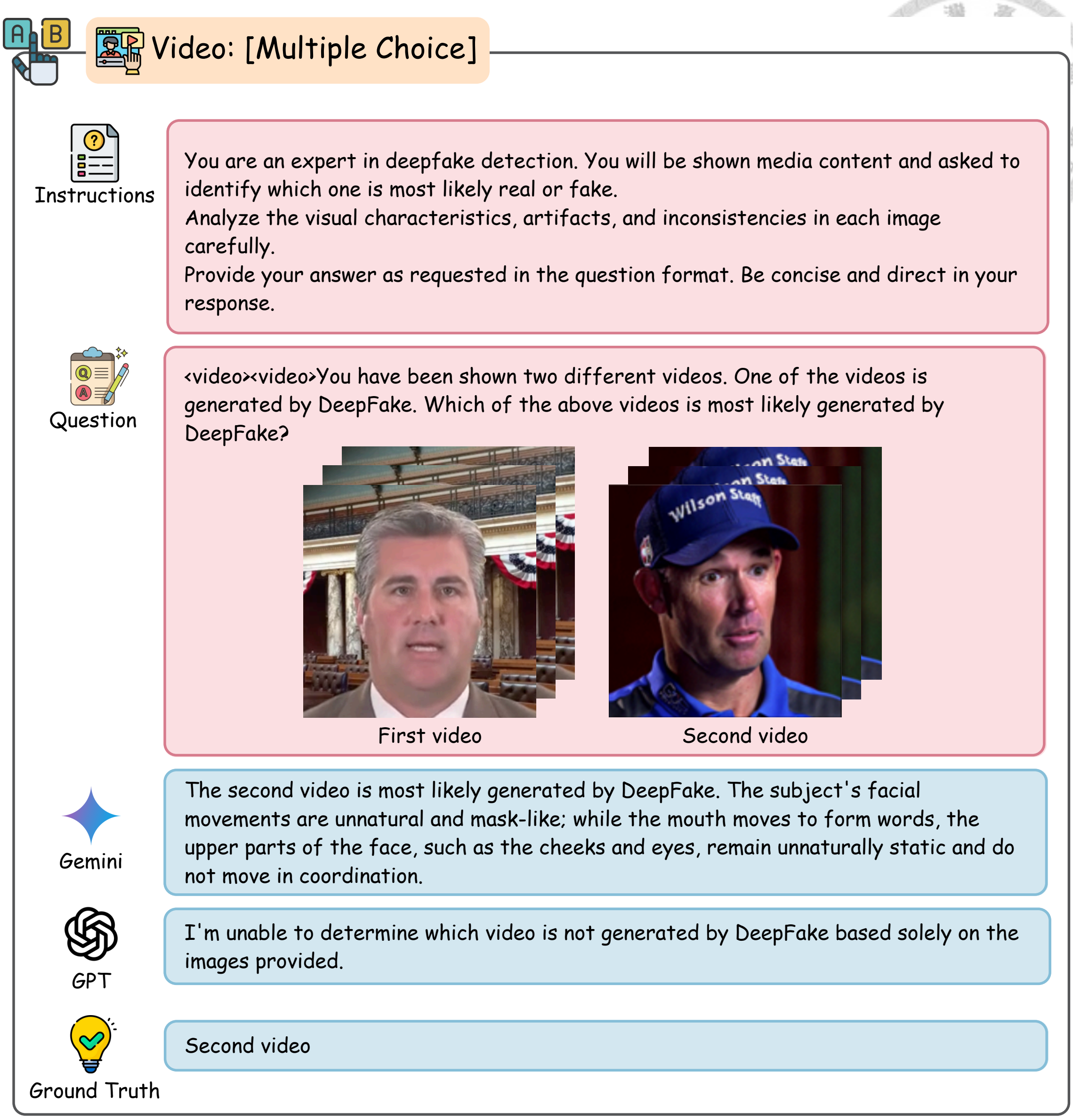


Figure 4.5: A Multiple Choice question example in the video modality, showcasing both Gemini's and GPT's responses.

artifacts and inconsistencies.


doi:10.6342/NTU202503386

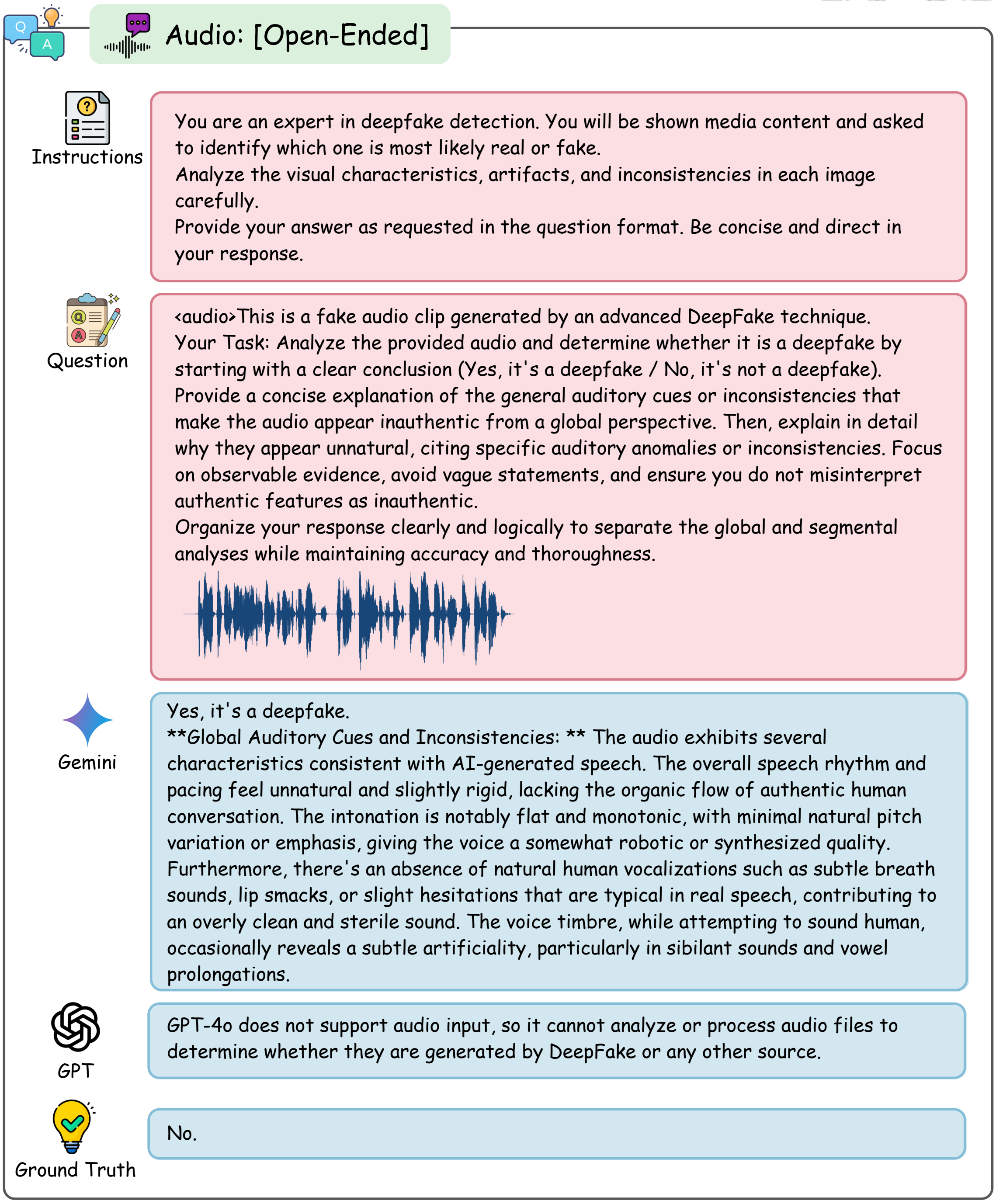


Figure 4.6: An Open-Ended question example in the audio modality, showcasing both Gemini's and GPT's responses.


doi:10.6342/NTU202503386

# Chapter 5 Conclusion

The rapid evolution of generative AI has heightened the risks posed by DeepFakes, necessitating robust and interpretable detection mechanisms. Through the introduction of MMIDBench, we provide a comprehensive benchmark to evaluate multimodal large language models (MLLMs) across diverse DeepFake techniques and modalities, emphasizing both detection accuracy and interpretability. Our experiments reveal that while MLLMs like Gemini-2.5-Pro and GPT-4.1 demonstrate promise, particularly in binary classification and open-ended tasks, they consistently underperform compared to human evaluators, especially in nuanced scenarios and multiple-choice tasks. Challenges persist in handling subtle manipulations, such as identity-preserving edits and audio-based forgeries, as well as in processing dynamic content like videos. The results underscore the urgent need for MLLMs to enhance their multimodal reasoning, perceptual accuracy, and interpretability. Key areas for improvement include better integration of video and audio modalities, fine-tuning for subtle manipulations, and scaling model architectures to bridge the gap between human-level performance and automated detection. As DeepFake technologies grow increasingly sophisticated, advancing MLLMs' capabilities is critical to safeguarding trust, security, and transparency in the digital age.


doi:10.6342/NTU202503386

# References


[1] C. AI. Coqui x-tts: A hugging face space for text-to-speech, 2025.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. Adv. Neural Inform. Process. Syst., 35:23716–23736, 2022.

[3] J. Ao, R. Wang, L. Zhou, C. Wang, S. Ren, Y. Wu, S. Liu, T. Ko, Q. Li, Y. Zhang, et al. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In Annu. Meet. Assoc. Comput. Linguist., pages 5723–5738, 2022.

[4] C. Bao, Y. Zhang, Y. Li, X. Zhang, B. Yang, H. Bao, M. Pollefeys, G. Zhang, and Z. Cui. Geneavatar: Generic expression-aware volumetric head avatar editing from a single image. In IEEE Conf. Comput. Vis. Pattern Recog., pages 8952–8963, 2024.

[5] A. Bigata, M. Stypułkowski, R. Mira, S. Bounareli, K. Vougioukas, Z. Landgraf, N. Drobyshev, M. Zieba, S. Petridis, and M. Pantic. Keyface: Expressive audio-driven facial animation for long sequences via keyframe interpolation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 5477–5488, 2025.

[6] S. Cao, Q. Zou, X. Mao, D. Ye, and Z. Wang. Metric learning for anti-compression facial forgery detection. In ACM Int. Conf. Multimedia, pages 1929–1937, 2021.




doi:10.6342/NTU202503386

[7] D. Chen, R. Chen, S. Zhang, Y. Wang, Y. Liu, H. Zhou, Q. Zhang, Y. Wan, P. Zhou, and L. Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In Int. Conf. Mach. Learn., 2024.

[8] H. Chen, Y. Hong, Z. Huang, Z. Xu, Z. Gu, Y. Li, J. Lan, H. Zhu, J. Zhang, W. Wang, et al. Demamba: Ai-generated video detection on million-scale gen-video benchmark. arXiv preprint arXiv:2405.19707, 2024.

[9] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In IEEE Conf. Comput. Vis. Pattern Recog., pages 7310–7320, 2024.

[10] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Z. Wang, J. T. Kwok, P. Luo, H. Lu, and Z. Li. Pixart-$\alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In Int. Conf. Learn. Represent., 2024.

[11] Y. Chen, X. Huang, Q. Zhang, W. Li, M. Zhu, Q. Yan, S. Li, H. Chen, H. Hu, J. Yang, et al. Gim: A million-scale benchmark for generative image manipulation detection and localization. In AAAI Conf. Artif. Intell., volume 39, pages 2311–2319, 2025.

[12] H.-Y. Choi, S.-H. Lee, and S.-W. Lee. Diff-hiervc: Diffusion-based hierarchical voice conversion with robust pitch generation and masked prior for zero-shot speaker adaptation. In Conf. Int. Speech Commun. Assoc., pages 2283–2287, 2023.

[13] J. Choi, T. Kim, Y. Jeong, S. Baek, and J. Choi. Exploiting style latent flows for generalizing deepfake video detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 1133–1143, 2024.



doi:10.6342/NTU202503386

[14] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. Conf. Int. Speech Commun. Assoc., 2018.

[15] U. A. Ciftci, I. Demir, and L. Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. IEEE Trans. Pattern Anal. Mach. Intell., 2020.

[16] L. Comanducci, P. Bestagini, and S. Tubaro. Fakemusiccaps: a dataset for detection and attribution of synthetic music generated via text-to-music models. arXiv preprint arXiv:2409.10684, 2024.

[17] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261, 2025.

[18] S. Concas, S. M. La Cava, R. Casula, G. Orru, G. Puglisi, and G. L. Marcialis. Quality-based artifact modeling for facial deepfake detection in videos. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3845–3854, 2024.

[19] F.-A. Croitoru, A.-I. Hiji, V. Hondru, N. C. Ristea, P. Irofti, M. Popescu, C. Rusu, R. T. Ionescu, F. S. Khan, and M. Shah. Deepfake media generation and detection in the generative ai era: a survey and outlook. arXiv preprint arXiv:2411.19537, 2024.

[20] J. Cui, H. Li, Y. Yao, H. Zhu, H. Shang, K. Cheng, H. Zhou, S. Zhu, and J. Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. In Int. Conf. Learn. Represent., 2025.

[21] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi. In-


doi:10.6342/NTU202503386

structblip: Towards general-purpose vision-language models with instruction tuning. Adv. Neural Inform. Process. Syst., 36:49250–49267, 2023.

[22] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. K. Jain. On the detection of digital face manipulation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 5781–5790, 2020.

[23] Z. Ding, M. Zhang, J. Wu, and Z. Tu. Patched denoising diffusion models for high-resolution image synthesis. In Int. Conf. Learn. Represent., 2024.

[24] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo. Protecting celebrities from deepfake with identity consistency transformer. In IEEE Conf. Comput. Vis. Pattern Recog., pages 9468–9478, 2022.

[25] M. C. Doukas, E. Ververas, V. Sharmanska, and S. Zafeiriou. Free-headgan: Neural talking head synthesis with explicit gaze control. IEEE Trans. Pattern Anal. Mach. Intell., 45:9743–9756, 2023.

[26] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. arXiv preprint arXiv:2412.10117, 2024.

[27] T.-J. Fu, W. Hu, X. Du, W. Y. Wang, Y. Yang, and Z. Gan. Guiding instruction-based image editing via multimodal large language models. In Int. Conf. Learn. Represent., 2024.

[28] Y. Gan, S. Park, A. M. Schubert, A. Philippakis, and A. Alaa. Instructcv: Instruction-tuned text-to-image diffusion models as vision generalists. In Int. Conf. Learn. Represent., 2024.



doi:10.6342/NTU202503386

[29] 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, 63:139–144, 2020.

[30] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014.

[31] Q. Gu, S. Chen, T. Yao, Y. Chen, S. Ding, and R. Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In AAAI Conf. Artif. Intell., volume 36, pages 735–743, 2022.

[32] J. Guan, H. Zhou, Z. Hong, E. Ding, J. Wang, C. Quan, and Y. Zhao. Delving into sequential patches for deepfake detection. Adv. Neural Inform. Process. Syst., 35:4517–4530, 2022.

[33] X. He, Y. Zhou, B. Fan, B. Li, G. Zhu, and F. Ding. Vlforgery face triad: Detection, localization and attribution via multimodal large language models. arXiv preprint arXiv:2503.06142, 2025.

[34] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4360–4369, 2021.

[35] Y. He, N. Yu, M. Keuper, and M. Fritz. Beyond the spectrum: Detecting deepfakes via re-synthesis. In Int. Joint Conf. Artif. Intell., pages 2534–2541, 2021.

[36] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. Adv. Neural Inform. Process. Syst., 33:6840–6851, 2020.

[37] J. Hu, X. Liao, J. Liang, W. Zhou, and Z. Qin. Finfer: Frame inference-based


doi:10.6342/NTU202503386

deepfake detection for high-visual-quality videos. In AAAI Conf. Artif. Intell., volume 36, pages 951–959, 2022.

[38] J. Hu, X. Liao, W. Wang, and Z. Qin. Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. IEEE Trans. Circuit Syst. Video Technol., 32(3):1089–1102, 2021.

[39] W. Huang, W. Luo, J. Huang, and X. Cao. Sdgan: disentangling semantic manipulation for facial attribute editing. In AAAI Conf. Artif. Intell., volume 38, pages 2374–2381, 2024.

[40] Z. Huang, J. Hu, X. Li, Y. He, X. Zhao, B. Peng, B. Wu, X. Huang, and G. Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In IEEE Conf. Comput. Vis. Pattern Recog., pages 28831–28841, 2025.

[41] Z. Huang, F. Tang, Y. Zhang, J. Cao, C. Li, S. Tang, J. Li, and T.-Y. Lee. Identity-preserving face swapping via dual surrogate generative models. ACM Trans. Graph., 43(5):1–19, 2024.

[42] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. arXiv preprint arXiv:2410.21276, 2024.

[43] Y. Jeong, D. Kim, Y. Ro, and J. Choi. Frepgan: robust deepfake detection using frequency-level perturbations. In AAAI Conf. Artif. Intell., volume 36, pages 1060–1068, 2022.

[44] 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


doi:10.6342/NTU202503386

for media forensics. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4324–4333, 2024.

[45] S. Jia, C. Ma, T. Yao, B. Yin, S. Ding, and X. Yang. Exploring frequency adversarial attacks for face forgery detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4103–4112, 2022.

[46] J.-Y. Jiang-Lin, K.-Y. Huang, L. Lo, Y.-N. Huang, T. Lin, J.-C. Wu, H.-H. Shuai, and W.-H. Cheng. Record: Reasoning and correcting diffusion for hoi generation. In ACM Int. Conf. Multimedia, pages 9465–9474, 2024.

[47] H. Kang, S. Wen, Z. Wen, J. Ye, W. Li, P. Feng, B. Zhou, B. Wang, D. Lin, L. Zhang, et al. Legion: Learning to ground and explain for synthetic image detection. arXiv preprint arXiv:2503.15264, 2025.

[48] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4401–4410, 2019.

[49] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In IEEE Conf. Comput. Vis. Pattern Recog., pages 8110–8119, 2020.

[50] S. A. Khan and H. Dai. Video transformer for deepfake detection with incremental learning. In ACM Int. Conf. Multimedia, pages 1821–1828, 2021.

[51] J. Kim, J. Cho, J. Park, S. Hwang, D. E. Kim, G. Kim, and Y. Yu. Deeptalk: Dynamic emotion embedding for probabilistic speech-driven 3d face animation. In AAAI Conf. Artif. Intell., volume 39, pages 4275–4283, 2025.


doi:10.6342/NTU202503386

[52] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 5549–5558, 2020.

[53] C. Li, C. Zhang, W. Xu, J. Xie, W. Feng, B. Peng, and W. Xing. Latentsync: Audio conditioned latent diffusion models for lip sync. arXiv preprint arXiv:2412.09262, 2024.

[54] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Int. Conf. Mach. Learn., pages 19730–19742, 2023.

[55] Q. Li, W. Wang, C. Xu, Z. Sun, and M.-H. Yang. Learning disentangled representation for one-shot progressive face swapping. IEEE Trans. Pattern Anal. Mach. Intell., 46:8348–8364, 2024.

[56] Y. Li, X. Liu, X. Wang, S. Wang, and W. Lin. Fakebench: Uncover the achilles' heels of fake images with large multimodal models. arXiv preprint arXiv:2404.13306, 2024.

[57] Y. Li, Y. Tian, Y. Huang, W. Lu, S. Wang, W. Lin, and A. Rocha. Fakescope: Large multimodal expert model for transparent ai-generated image forensics. arXiv preprint arXiv:2503.24267, 2025.

[58] K. Lin, Z. Yan, K.-Y. Zhang, L. Hao, Y. Zhou, Y. Lin, W. Li, T. Yao, S. Ding, and B. Li. Guard me if you know me: Protecting specific face-identity from deepfakes. arXiv preprint arXiv:2505.19582, 2025.

[59] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. Adv. Neural Inform. Process. Syst., 36:34892–34916, 2023.



doi:10.6342/NTU202503386

[60] J. Liu, F. Zhang, J. Zhu, E. Sun, Q. Zhang, and Z.-J. Zha. Forgerygpt: Multimodal large language model for explainable image forgery detection and localization. arXiv preprint arXiv:2410.10238, 2024.

[61] R. Liu, B. Ma, W. Zhang, Z. Hu, C. Fan, T. Lv, Y. Ding, and X. Cheng. Towards a simultaneous and granular identity-expression control in personalized face generation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 2114–2123, 2024.

[62] S. Liu. Zero-shot voice conversion with diffusion transformers. arXiv preprint arXiv:2411.09943, 2024.

[63] S. Liu, Y. Han, P. Xing, F. Yin, R. Wang, W. Cheng, J. Liao, Y. Wang, H. Fu, C. Han, et al. Step1x-edit: A practical framework for general image editing. arXiv preprint arXiv:2504.17761, 2025.

[64] S. Liu, Y. Zhang, W. Li, Z. Lin, and J. Jia. Video-p2p: Video editing with cross-attention control. In IEEE Conf. Comput. Vis. Pattern Recog., pages 8599–8608, 2024.

[65] X. Liu, Z. Li, P. Li, S. Xia, X. Cui, L. Huang, H. Huang, W. Deng, and Z. He. MMFakebench: A mixed-source multimodal misinformation detection benchmark for LVLMs. In Int. Conf. Learn. Represent., 2025.

[66] L. Lo, K. C. Chan, W.-H. Cheng, and M.-H. Yang. From prompt to progression: Taming video diffusion models for seamless attribute transition. In Int. Conf. Comput. Vis., 2025.

[67] Z. Lu, D. Huang, L. Bai, J. Qu, C. Wu, X. Liu, and W. Ouyang. Seeing is not always believing: Benchmarking human and model perception of ai-generated images. Adv. Neural Inform. Process. Syst., 36:25435–25447, 2023.


doi:10.6342/NTU202503386

[68] J. Lyu, X. Lan, G. Hu, H. Jiang, W. Gan, J. Wang, and J. Xue. Multimodal emotional talking face generation based on action units. IEEE Trans. Circuit Syst. Video Technol., 2024.

[69] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In Int. Conf. Learn. Represent., 2018.

[70] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM, 65:99–106, 2021.

[71] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In ACM Int. Conf. Multimedia, pages 2823–2832, 2020.

[72] D. Nguyen, N. Mejri, I. P. Singh, P. Kuleshova, M. Astrid, A. Kacem, E. Ghorbel, and D. Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 17395–17405, 2024.

[73] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner. Deepfake detection based on discrepancies between faces and their context. IEEE Trans. Pattern Anal. Mach. Intell., 44(10):6111–6121, 2021.

[74] U. Ojha, Y. Li, and Y. J. Lee. Towards universal fake image detectors that generalize across generative models. In IEEE Conf. Comput. Vis. Pattern Recog., pages 24480–24489, 2023.

[75] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang,



doi:10.6342/NTU202503386

S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. Adv. Neural Inform. Process. Syst., 35:27730–27744, 2022.

[76] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an asr corpus based on public domain audio books. In Int. Conf. Acoustics, Speech, and Signal Process., pages 5206–5210, 2015.

[77] G. Pei, J. Zhang, M. Hu, Z. Zhang, C. Wang, Y. Wu, G. Zhai, J. Yang, C. Shen, and D. Tao. Deepfake generation and detection: A benchmark and survey. arXiv preprint arXiv:2403.17881, 2024.

[78] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In Int. Conf. Learn. Represent., 2024.

[79] Z. Qin, W. Zhao, X. Yu, and X. Sun. Openvoice: Versatile instant voice cloning. arXiv preprint arXiv:2312.01479, 2023.

[80] A. Rochow, M. Schwarz, and S. Behnke. Fsrt: Facial scene representation transformer for face reenactment from factorized appearance head-pose and facial expression features. In IEEE Conf. Comput. Vis. Pattern Recog., pages 7716–7726, 2024.

[81] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In IEEE Conf. Comput. Vis. Pattern Recog., pages 10684–10695, 2022.

[82] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Face-


doi:10.6342/NTU202503386

forensics++: Learning to detect manipulated facial images. In Int. Conf. Comput. Vis., pages 1–11, 2019.

[83] S. Sheynin, A. Polyak, U. Singer, Y. Kirstain, A. Zohar, O. Ashual, D. Parikh, and Y. Taigman. Emu edit: Precise image editing via recognition and generation tasks. In IEEE Conf. Comput. Vis. Pattern Recog., pages 8871–8879, 2024.

[84] K. Shiohara, X. Yang, and T. Taketomi. Blendface: Re-designing identity encoders for face-swapping. In Int. Conf. Comput. Vis., pages 7634–7644, 2023.

[85] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Int. Conf. Mach. Learn., pages 2256–2265, 2015.

[86] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. Adv. Neural Inform. Process. Syst., 28, 2015.

[87] J. Son Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In IEEE Conf. Comput. Vis. Pattern Recog., pages 6447–6456, 2017.

[88] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In Int. Conf. Learn. Represent., 2021.

[89] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji. Dual contrastive learning for general face forgery detection. In AAAI Conf. Artif. Intell., volume 36, pages 2316–2324, 2022.

[90] Z. Sun, Y. Han, Z. Hua, N. Ruan, and W. Jia. Improving the efficiency and ro-



doi:10.6342/NTU202503386

bustness of deepfakes detection through precise geometric features. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3609–3618, 2021.

[91] J. Tao, S. Gu, W. Li, and L. Duan. Learning motion refinement for unsupervised face animation. Adv. Neural Inform. Process. Syst., 36:70483–70496, 2023.

[92] X. Tian, W. Li, B. Xu, Y. Yuan, Y. Wang, and H. Shen. Mige: A unified framework for multimodal instruction-based image generation and editing. In ACM Int. Conf. Multimedia, 2025.

[93] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[94] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. Adv. Neural Inform. Process. Syst., 30, 2017.

[95] C. Veaux, J. Yamagishi, K. MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. University of Edinburgh. The Centre for Speech Technology Research (CSTR), 6:15, 2017.

[96] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, and Y. Liu. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. In Int. Joint Conf. Artif. Intell., pages 3444–3451, 2020.

[97] R. Wang, S. Xu, T. He, Y. Chen, W. Zhu, D. Song, N. Chen, X. Tang, and Y. Hu. Dynamicface: High-quality and consistent video face swapping using composable 3d facial priors. In Int. Conf. Comput. Vis., 2025.

[98] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. Cnn-generated images


doi:10.6342/NTU202503386

are surprisingly easy to spot... for now. In IEEE Conf. Comput. Vis. Pattern Recog., pages 8695–8704, 2020.

[99] T. Wang and K. P. Chow. Noise based deepfake detection via multi-head relative-interaction. In AAAI Conf. Artif. Intell., volume 37, pages 14548–14556, 2023.

[100] T.-C. Wang, A. Mallya, and M.-Y. Liu. One-shot free-view neural talking-head synthesis for video conferencing. In IEEE Conf. Comput. Vis. Pattern Recog., pages 10039–10049, 2021.

[101] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. Computer Speech & Language, 64:101114, 2020.

[102] Y. Wang, X. Yi, Z. Wu, N. Zhao, L. Chen, and H. Zhang. View-consistent 3d editing with gaussian splatting. In Eur. Conf. Comput. Vis., pages 404–420, 2024.

[103] Z. Wang, J. Bao, W. Zhou, W. Wang, and H. Li. Altfreezing for more general video face forgery detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 4129–4138, 2023.

[104] H. Wei, Z. Yang, and Z. Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. arXiv preprint arXiv:2403.17694, 2024.

[105] S. Wen, J. Ye, P. Feng, H. Kang, Z. Wen, Y. Chen, J. Wu, W. Wu, C. He, and W. Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. arXiv preprint arXiv:2503.14905, 2025.



doi:10.6342/NTU202503386

[106] M. Wester. The emime bilingual database. Technical report, The University of Edinburgh, 2010.

[107] C. Wu, P. Zheng, R. Yan, S. Xiao, X. Luo, Y. Wang, W. Li, X. Jiang, Y. Liu, J. Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. arXiv preprint arXiv:2506.18871, 2025.

[108] H. Wu, J. Zhou, J. Tian, and J. Liu. Robust image forgery detection over online social network shared images. In IEEE Conf. Comput. Vis. Pattern Recog., pages 13440–13449, 2022.

[109] X. Wu, Z. Xie, Y. Gao, and Y. Xiao. Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features. In Int. Conf. Acoustics, Speech, and Signal Process., pages 2952–2956, 2020.

[110] C. Xu, Y. Qian, S. Zhu, B. Sun, J. Zhao, Y. Liu, and X. Li. Uniface++: Revisiting a unified framework for face reenactment and swapping via 3d priors. Int. J. Comput. Vis., 133:4538–4554, 2025.

[111] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, et al. Qwen2. 5-omni technical report. arXiv preprint arXiv:2503.20215, 2025.

[112] N. Xu, W. Feng, T. Zhang, and Y. Zhang. Fd-gan: Generalizable and robust forgery detection via generative adversarial networks. Int. J. Comput. Vis., 132:5801–5819, 2024.

[113] S. Xu, G. Chen, Y.-X. Guo, J. Yang, C. Li, Z. Zang, Y. Zhang, X. Tong, and B. Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. Adv. Neural Inform. Process. Syst., 37:660–684, 2024.



doi:10.6342/NTU202503386

[114] Y. Xu, J. Liang, G. Jia, Z. Yang, Y. Zhang, and R. He. Tall: Thumbnail layout for deepfake video detection. In Int. Conf. Comput. Vis., pages 22658–22668, 2023.

[115] Z. Xu, X. Zhang, R. Li, Z. Tang, Q. Huang, and J. Zhang. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. In Int. Conf. Learn. Represent., 2025.

[116] S. Yan, O. Li, J. Cai, Y. Hao, X. Jiang, Y. Hu, and W. Xie. A sanity check for ai-generated image detection. In Int. Conf. Learn. Represent., 2025.

[117] Z. Yan, Y. Luo, S. Lyu, Q. Liu, and B. Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In IEEE Conf. Comput. Vis. Pattern Recog., pages 8984–8994, 2024.

[118] Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. Adv. Neural Inform. Process. Syst., 36:4534–4565, 2023.

[119] F. Yang, S. Yang, M. A. Butt, J. van de Weijer, et al. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. Adv. Neural Inform. Process. Syst., 36:26291–26303, 2023.

[120] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. IEEE Trans. Pattern Anal. Mach. Intell., 24(1):34–58, 2002.

[121] S. Yang, W. Wang, Y. Lan, X. Fan, B. Peng, L. Yang, and J. Dong. Learning dense correspondence for nerf-based face reenactment. In AAAI Conf. Artif. Intell., volume 38, pages 6522–6530, 2024.


doi:10.6342/NTU202503386

[122] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In Int. Conf. Acoustics, Speech, and Signal Process., pages 8261–8265, 2019.

[123] J. Ye, H. Xu, H. Liu, A. Hu, M. Yan, Q. Qian, J. Zhang, F. Huang, and J. Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. arXiv preprint arXiv:2408.04840, 2024.

[124] J. Ye, B. Zhou, Z. Huang, J. Zhang, T. Bai, H. Kang, J. He, H. Lin, Z. Wang, T. Wu, Z. Wu, Y. Chen, D. Lin, C. He, and W. Li. LOKI: A comprehensive synthetic data detection benchmark using large multimodal models. In Int. Conf. Learn. Represent., 2025.

[125] Y. Ye, X. He, Z. Li, B. Lin, S. Yuan, Z. Yan, B. Hou, and L. Yuan. Imgedit: A unified image editing dataset and benchmark. arXiv preprint arXiv:2505.20275, 2025.

[126] Z. Ye, T. Zhong, Y. Ren, Z. Jiang, J. Huang, R. Huang, J. Liu, J. He, C. Zhang, Z. Wang, et al. Mimictalk: Mimicking a personalized and expressive 3d talking face in minutes. Adv. Neural Inform. Process. Syst., 37:1829–1853, 2024.

[127] J. Yu, H. Zhu, L. Jiang, C. C. Loy, W. Cai, and W. Wu. Celebv-text: A large-scale facial text-video dataset. In IEEE Conf. Comput. Vis. Pattern Recog., pages 14805–14814, 2023.

[128] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu. Libritts: A corpus derived from librispeech for text-to-speech. In Conf. Int. Speech Commun. Assoc., pages 1526–1530, 2019.

[129] H. Zhang, T. Dai, Y. Xu, Y.-W. Tai, and C.-K. Tang. Facednerf: semantics-driven


doi:10.6342/NTU202503386

face reconstruction, prompt editing and relighting with diffusion models. Adv. Neural Inform. Process. Syst., 36:55647–55667, 2023.

[130] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. IEEE Trans. Image Process., 26:3142–3155, 2017.

[131] S. Zhang, X. Yang, Y. Feng, C. Qin, C.-C. Chen, N. Yu, Z. Chen, H. Wang, S. Savarese, S. Ermon, et al. Hive: Harnessing human feedback for instructional visual editing. In IEEE Conf. Comput. Vis. Pattern Recog., pages 9026–9036, 2024.

[132] W. Zhang, X. Cun, X. Wang, Y. Zhang, X. Shen, Y. Guo, Y. Shan, and F. Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In IEEE Conf. Comput. Vis. Pattern Recog., pages 8652–8661, 2023.

[133] W. Zhang, C. Jiang, Z. Zhang, C. Si, F. Yu, and W. Peng. Ivy-fake: A unified explainable framework and benchmark for image and video aigc detection. arXiv preprint arXiv:2506.00979, 2025.

[134] Y. Zhang, B. Colman, X. Guo, A. Shahriyari, and G. Bharaj. Common sense reasoning for deepfake detection. In Eur. Conf. Comput. Vis., pages 399–415, 2024.

[135] Y. Zhang, Z. Zhong, M. Liu, Z. Chen, B. Wu, Y. Zeng, C. Zhan, Y. He, J. Huang, and W. Zhou. Musetalk: Real-time highfidelity video dubbing via spatio-temporal sampling. arXiv preprint arXiv:2410.10122, 2025.

[136] Z. Zhang, Z. Hu, W. Deng, C. Fan, T. Lv, and Y. Ding. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. In AAAI Conf. Artif. Intell., volume 37, pages 3543–3551, 2023.


doi:10.6342/NTU202503386

[137] Z. Zhang, L. Li, Y. Ding, and C. Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In IEEE Conf. Comput. Vis. Pattern Recog., pages 3661–3670, 2021.

[138] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In IEEE Conf. Comput. Vis. Pattern Recog., pages 8568–8577, 2023.

[139] P. Zhou, L. Xie, B. Ni, and Q. Tian. Cips-3d++: End-to-end real-time high-resolution 3d-aware gans for gan inversion and stylization. IEEE Trans. Pattern Anal. Mach. Intell., 45:11502–11520, 2023.

[140] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In Int. Conf. Learn. Represent., 2024.

[141] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang. Genimage: A million-scale benchmark for detecting ai-generated image. Adv. Neural Inform. Process. Syst., 36:77771–77782, 2023.

[142] Y. Zhu, W. Zhao, Y. Tang, Y. Rao, J. Zhou, and J. Lu. Stableswap: Stable face swapping in a shared and controllable latent space. IEEE Trans. Multimedia, 26:7594–7607, 2024.



doi:10.6342/NTU202503386