APSIPA Transactions on Signal and Information Processing, 2025, 14, e202

Original Paper

Navigating Real and Fake in the Era of Advanced Generative AI

Huy H. Nguyen[1*], Siyun Liang[2], Junichi Yamagishi[1] and Isao Echizen[1, 3]

[1] *National Institute of Informatics, Tokyo, Japan*
[2] *Technical University of Munich, Munich, Germany*
[3] *The University of Tokyo, Tokyo, Japan*

ABSTRACT

In this era of advanced generative artificial intelligence (AI), in which machine-generated content coexists with human-created content, the question of authenticity extends beyond the binary "real or fake." Media forensics must evolve to encompass three crucial dimensions. **Provenance**: was the content created by a person, AI, or a combination of both? **Intention**: was the content created with a specific purpose, such as to inform, entertain, deceive, or manipulate? **Context**: how is the content being presented, used, and interpreted within its particular context? As AI-generated content becomes increasingly prevalent, the focus should shift towards ensuring transparency, ethical use, and accountability in content creation and dissemination, regardless of its origin.

In this paper, we present a comprehensive countermeasure framework designed to address a broad spectrum of attacks related to generative AI, where merely distinguishing between "real" and "fake" content is no longer adequate. To highlight the necessity of this new perspective, we present two case studies that demonstrate the feasibility and effectiveness of our proposed framework. By implementing such a framework, we can more effectively navigate the

*Corresponding author: huyhnguyen.work@gmail.com.

Received 01 December 2024; revised 15 March 2025; accepted 08 April 2025
ISSN 2048-7703; DOI 10.1561/116.20240091

challenges posed by advanced generative AI while harnessing the opportunities it presents.

Keywords: AI-generated content, deepfake, real or fake, generative AI, AI-powered framework, countermeasures, provenance, intention, context

1 Introduction

Over the past decade, generative AI has experienced significant advancements. The introduction of generative adversarial networks [6] and the subsequent development of denoising diffusion probabilistic models [10] have revolutionized the field. These innovations enable AI to generate highly realistic multimedia content, including images, videos, speeches, and text, with and without additional information, providing greater control and usability. As a result, generative AI has attracted substantial investments in research and development, paving the way for various commercial applications such as conversational agents, design assistants, content creation tools, and entertainment services.

However, concerns about information security have also escalated. The rise of generative AI applications, such as deepfakes [18], which leverage deep learning techniques to synthesize or manipulate media content and often incorporate human biometric traits, poses a significant threat to the authenticity and trustworthiness of digital media. To address this threat, researchers in the field of media forensics have been actively exploring ways to detect deepfakes, particularly manipulated 2D images, videos [1], and speeches [8]. Recent advancements, such as the emergence of large language models (LLMs) and large vision language models, such as ChatGPT[1] (proprietary) or Llama [4] (open-source), have also spurred the development of methods for detecting AI-generated text [14, 20].

Most deepfake detection approaches can be characterized as binary classification problems, often relying on machine learning, particularly deep learning technologies. To train detectors to distinguish between real and fake content, researchers assemble large-scale datasets containing both genuine and fake content. However, the increasing usage of generative AI to enhance content has highlighted that differentiating between "real" and "fake" content is no longer sufficient. In particular, advanced media processing techniques can be used both by attackers creating deepfake content and by benign users for legitimate content creation, as shown in Figure 1. For instance, NVIDIA Maxine,[2]

[1] https://openai.com/blog/chatgpt
[2] https://developer.nvidia.com/maxine

Eye contact adjustment with NVIDIA Maxine.

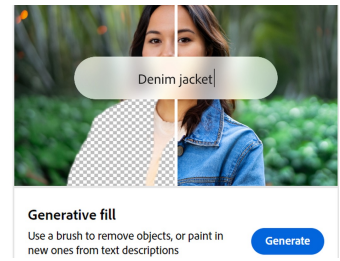
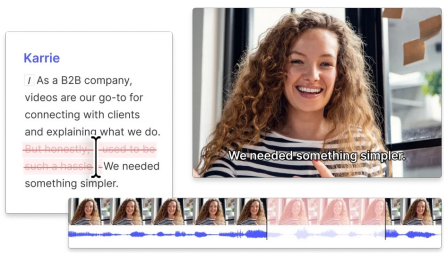
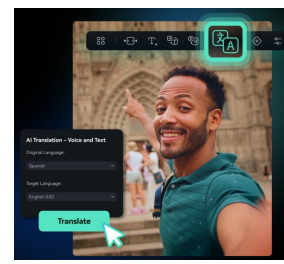


Image editing with Adobe Firefly.



Video editing by editing the script with Descript.



Video audio translation with Filmora.

Figure 1: Examples of AI-powered applications used by benign users for legitimate content creation. Such applications can be used by adversaries to create malicious content. Images were obtained from relevant company web pages or product introduction websites.

an AI-powered real-time video communication system, offers features such as maintaining eye contact and real-time speech translation. While these capabilities enhance video communication, they overlap with techniques like facial reenactment and speech synthesis commonly used by attackers to produce deepfake videos and audio. Another example is the video editing functionality offered by Descript,[3] which enables users to modify video content by editing its script. Although this feature is intended to streamline video production, attackers can exploit it to create audio-visual deepfakes. Similarly, LLMs used as writing assistants can seamlessly blend human-written and AI-generated text. This behavior is similar to the process of biased content generation [7], further complicating the task of identifying authentic versus manipulated content.

These trends present challenges for traditional data-driven deepfake detection approaches and raise **two key research problems**. **First**, *the boundary between real and fake is becoming increasingly unclear, making it difficult to assign simple binary labels to content.* For instance, determining whether a caller using the NVIDIA Maxine communication system has malicious intent

[3] https://www.descript.com

is challenging. Standard deepfake detectors might classify benign videos as fake because they were manipulated by AI. However, redesigning the training data to include such videos as real will cause confusion in the model. Moreover, if an attacker compromises NVIDIA Maxine, a detector trained on the redesigned dataset could classify malicious videos as real because their characteristics are similar to those of the "benign enhanced" videos.

Second, *determining the provenance, intention, and context in which content was produced and presented is crucial for accurate detection.* Media content can be created by people, generated by AI, or collaboratively produced by both. Understanding the intention behind content creation and assessing whether the use of generative AI was legitimate or malicious is essential for context-aware detection. Unfortunately, most current applications and systems are not designed to support such understanding. Initial attempts, such watermarking text generated by LLMs for content authentication [11, 3], are simply a first step. There remains a significant need for further research and development, especially in the areas of multi-domain and international standardization, to comprehensively tackle this problem. The field of media forensics must undergo a revolution to keep pace with the rapid development and widespread application of generative AI.

To address these challenges, we present a generalized AI-powered framework encompassing the design of most modern AI-driven systems and platforms. Using this framework, we demonstrate that traditional deepfake detection approaches, which operate primarily at the end-user level, are insufficient for distinguishing malicious manipulations from legitimate AI-assisted enhancements. We systematically identify critical points within the framework that attackers can exploit to manipulate hybrid human-AI multimedia content and suggest corresponding defense mechanisms to form a comprehensive countermeasure framework. To further illustrate the practical implications of our approach, we present two case studies focusing on different AI-powered applications: an interactive communication framework and a non-interactive content editing framework. These use cases were inspired by widely used AI systems such as NVIDIA Maxine, Descript, and Filmora.[4] These examples highlight the need to adopt a broader approach to AI-driven content authentication that goes beyond conventional forensic techniques and emphasizes provenance, intention, and context.

The remainder of the paper is structured as follows: Section 2 explores the evolving landscape of deepfake detection, contrasting traditional approaches with the challenges posed by generative AI. Section 3 introduces a generalized AI-powered framework that captures the architecture of modern AI-driven systems and serves as the foundation for our security analysis. Section 4 systematically examines potential deepfake attack vectors within this framework,

[4] https://filmora.wondershare.net

while Section 5 outlines a comprehensive countermeasure strategy to mitigate these threats. The practical implications of our approach are demonstrated by the two case studies in Sections 6 and 7: one on an interactive AI-powered communication system and the other on a non-interactive AI-driven content editing platform. These real-world examples highlight the need to move beyond binary deepfake detection and embrace a security paradigm that considers provenance, intention, and context. Finally, we conclude with a summary of the key points in Section 8.

2 Benign Users and Adversaries in the Era of Generative AI

In this section, we explore the key differences between the settings commonly assumed in traditional deepfake detection research and the proposed framework tailored for the generative AI era. The key differences are summarized in Figure 2.

2.1 Traditional Deepfake Setting

In the traditional deepfake detection setting, which was used in most previous studies, benign users are assumed to not use generative AI to enhance or modify their content. Adversaries, on the other hand, either create deepfakes from scratch—synthesizing entire images, videos, speech, or text—or manipulate existing content obtained from victims using adversary-controlled driving information. *The method of acquiring such media is often considered irrelevant and is typically assumed to involve publicly available sources, such as the Internet.*

Deepfake detection in this setting generally involves *binary classification*, the goal of which is to predict the likelihood that a given media file is a deepfake. In some cases, multi-class classification is used to identify the specific type of deepfake. Additionally, *segmentation* techniques may be used to localize the spatial regions or temporal segments that have been manipulated.

2.2 Generative AI Era Setting

In the generative AI era setting, unlike the traditional setting, *benign users may utilize an AI-powered system to assist in content creation*, resulting in content that *blends human and machine-generated elements.* Such systems use techniques similar to those used for generating deepfakes. Due to the high computational power required, such systems are typically offered as public cloud services, making them accessible to general users. In contrast, adversaries often deploy deepfake methods locally or privately, primarily because

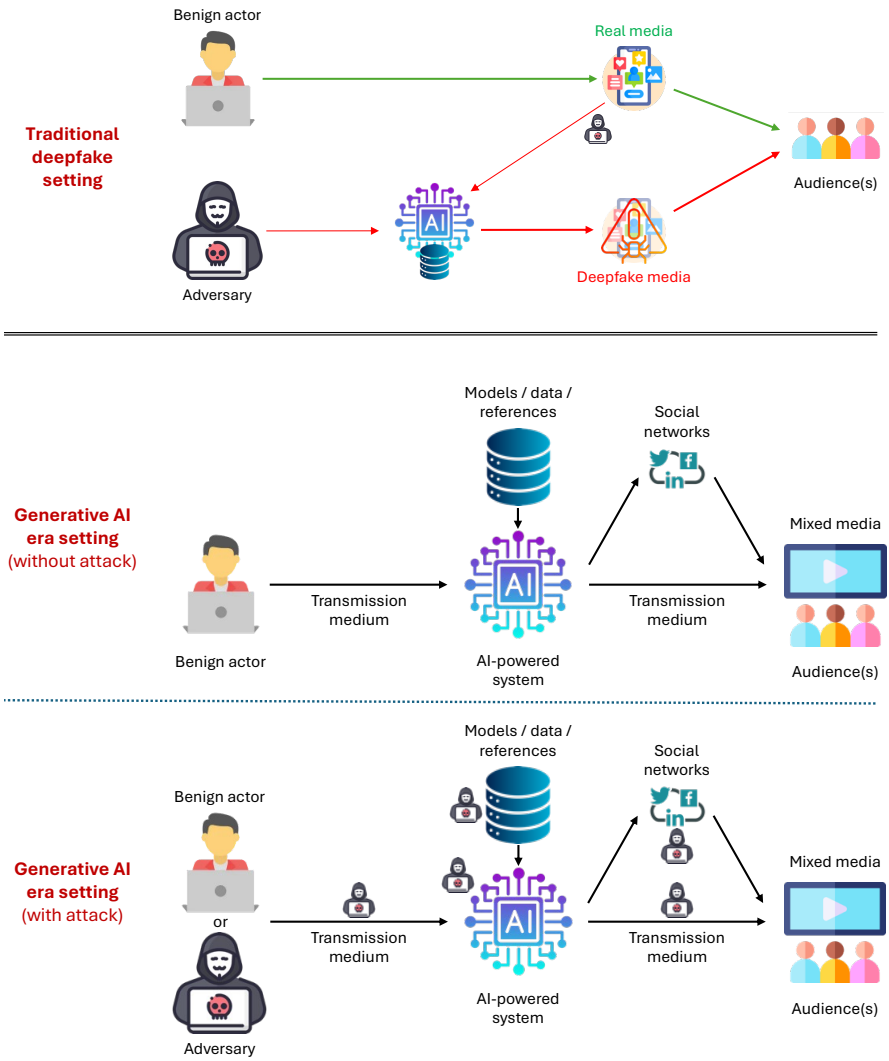


Figure 2: Benign users and adversaries before and during the generative AI setting era. In the **traditional deepfake setting**, benign users do not use generative AI to enhance or alter their content, whereas adversaries do. The approach used to acquire the victim's media is typically not a focal point. In the **generative AI era setting**, both parties utilize generative AI for creating or modifying media content. Benign users typically rely on an **online** generative AI system, while adversaries generally prefer a **local** system for deepfake generation. We further divide the generative AI era setting into two cases: with and without attacks.

most public AI-powered services prohibit malicious use and present risks of exposing the adversaries' activities or identities.

In this setting, *the methods for obtaining a victim's media and disseminating malicious AI content take on greater significance.* Beyond creating original media content or manipulating stored or shared content, adversaries may act as intermediaries (*e.g.*, man-in-the-middle attacks) or target AI-powered systems directly by compromising models, datasets, or system decisions. The situation becomes even more complex when adversaries infiltrate the content creation process of benign users using an AI-powered system. In such cases, the resulting media integrates benign hybrid content with malicious elements introduced by adversaries.

Figure 3 illustrates an example of varying levels of machine involvement in human-created content within the domain of 3D facial modeling [5]. Creating a 3D facial model requires a combination of facial shape and texture that can later be used to render 2D or 3D facial images or videos. These components can be derived from a real person using precise scanners or through 3D reconstruction techniques [15]. Alternatively, they can be reconstructed from entirely synthesized facial images or generated directly by a 3D generative model. Subsequently, the shape and texture can be further manipulated by benign users using an AI-powered system or compromised by adversaries. In the latter case, the result is complex hybrid content involving three entities: the original creator, the AI system, and the adversary.

As a result, binary classification approaches used in traditional deepfake detection are inadequate for this generative AI era. Instead, there is a need to analyze the *provenance, intention, and context* of media production to distinguish benign from malicious usage.

3 A Generalized AI-powered Framework

Various AI-powered frameworks have been tailored for specific applications, such as communication, content creation, and extended reality. These frameworks can include various components, utilize cloud computing or not, and cater to diverse user bases. We present a general framework for comprehending these variations; it is visually depicted in Figure 4. Some components are optional depending on the application.

In this generalized AI-powered framework, the actor is a person creating original content. The audience can be either the same individual or a different entity. Since we live in an analog world, the framework requires at least one analog-to-digital converter (ADC) and at least one digital-to-analog converter (DAC). The ADC can be a camera, light detection and ranging (LiDAR) system, microphone, keyboard, or digital sketch board. The DAC can be a monitor, projector, speaker, or headphones. In the case of multimedia content,

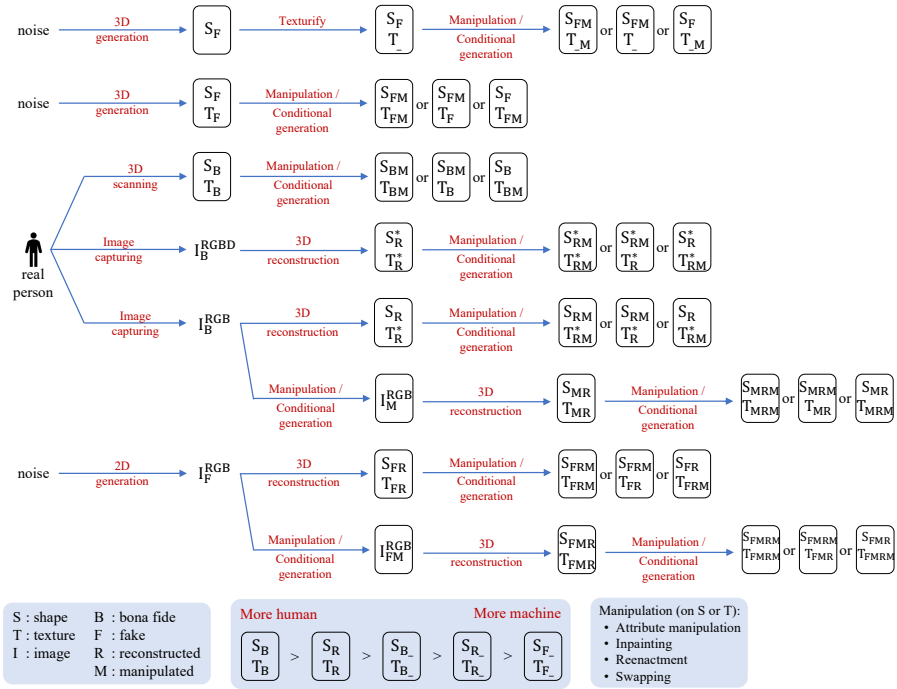


Figure 3: Levels of machine involvement in 3D face modeling in generative AI era setting. A 3D model can be created or reconstructed on the basis of a real person or an AI-generated individual. Alternatively, the 3D model can be synthesized from scratch. Once the shape and texture of the face are established, further manipulations can be performed by either benign users or adversaries.

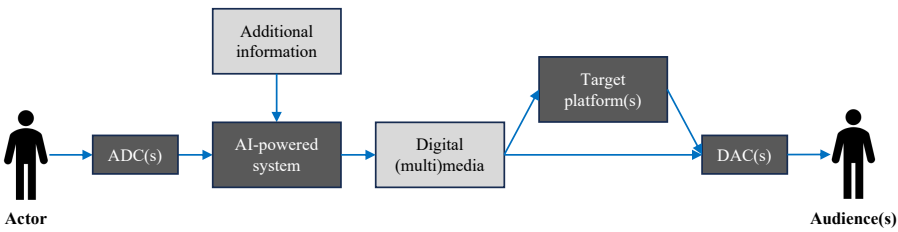


Figure 4: A generalized framework for AI-powered system. For the case of communication systems like Maxine, call participants play the roles of both speaker and listener. Depending on the applications, some components are optional.

multiple ADCs and DACs are required. The additional information could be controlling information (*e.g.*, destination language for machine translation or viewer's pose for communication) or other relevant data. The target platform

which host the multimedia content could be a social network or an online platform such as a learning management system.

An interactive example of the generalized AI-powered framework is NVIDIA's Maxine. In this example, the actor is the speaker, and the audience consists of the listener(s). Beyond basic video calling functionality, the system incorporates advanced features such as voice conversion and lip-syncing for real-time language translation, creating the illusion that the speaker is communicating in the listener's native language. Additionally, the system can enhance eye contact by adjusting the speaker's gaze on the basis of the positions of both the speaker and listener. The additional information indicated in Figure 4 includes the listener's language and RGB-D videos recorded by RGB and depth cameras. The enhanced audio-visual content is directly delivered to the listener(s) without passing through an external target platform.

A non-interactive example of the generalized AI-powered framework is Descript. In this example, the actor is the video content creator. Descript first generates the transcript for the video captured by the user. The user can then edit the video by editing the script. Descript edits the visual and audio parts of the video to match the edited transcript. Regarding the target platform indicated in Figure 4, the edited video can be shared on social networks or other online platforms. It can be also presented directly to the audience(s).

4 Deepfake Attacks on the AI-powered Frameworks

Similar to other cyber systems, AI-powered systems are vulnerable to exploitation by adversaries who generate or manipulate multimedia content, thereby complicating the verification of authenticity. This paper focuses on attacks that inject adversarial information, resulting in alterations to the final media content. Adversaries may include the actor, a third party compromising the system's components, or a man-in-the-middle operating at various points in the transmission chain—between the actor, the system modules, and the audience(s), or between the modules themselves. Inspired by Ratha *et al.* [17], we model the potential attack vectors as depicted in Figure 5. In detail, adversaries can exploit the system in various ways, including:

- **Performing a presentation attack [13] on the ADC(s)** ① using deepfake material. Unlike traditional presentation attacks that rely on pre-captured real images, videos, or audio, the deepfake material here may involve entirely synthesized media or manipulated existing media.
- **Intercepting the ADC(s)** ② **or DAC(s)** ⑪ to inject deepfake material. For ②, the actor may perform the interception to avoid pre-

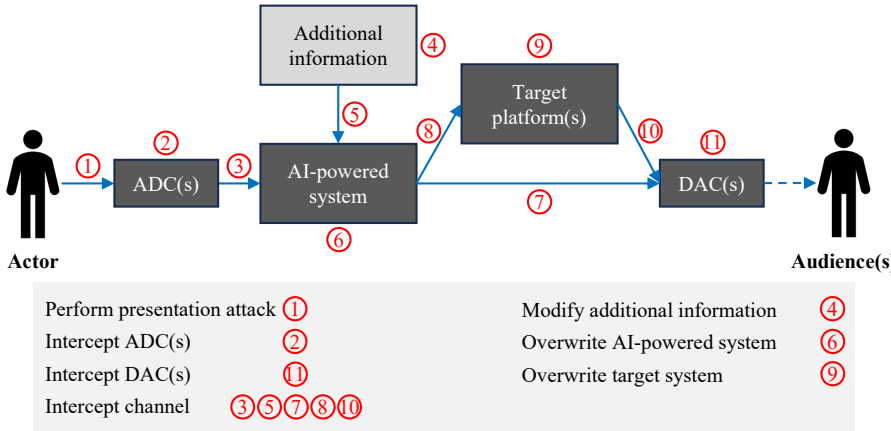


Figure 5: Critical points in AI-powered framework susceptible to deepfake-related attacks.

sentation attack artifacts from ①, or a separate adversary might do it.

- **Intercepting communication channels**, i.e., a man-in-the-middle attack, between devices, systems, or system modules at various points such as ③, ⑤, ⑦, ⑧, and ⑩, enabling the adversary to manipulate transmitted data.
- **Manipulating the additional information** ④ provided to the AI-powered system. This typically involves altering control information derived from the audience and using it to adapt the actor's media to the audience's preferences.
- **Tampering with components of the AI-powered system** ⑥, such as its models, training data, workflows, and decision modules. Under certain circumstances, the AI-powered system itself may act as an adversary.
- **Overwriting the target platform** ⑨ by replacing benign user content with deepfake content. Under certain circumstances, the target platform itself may function as an adversary.

This structured view elucidates the diverse attack surfaces in AI-powered systems in the era of generative AI, particularly with regard to deepfake-related attacks. It highlights the need for robust countermeasures, which will be discussed in the next section.

5 Deepfake Countermeasures in the Era of Generative AI

Given the generalized AI-powered framework introduced in Section 3 and the attack scenarios detailed in Section 4, it is evident that traditional deepfake detection methods, which are typically deployed on the audience's side, are inadequate for addressing the complexities of the generative AI era. To effectively mitigate deepfake threats, we propose using a comprehensive approach that integrates various countermeasures at multiple points within the generalized AI framework. These measures aim to establish the provenance, intent, and context of media production and presentation while countering diverse attack vectors and include, but are not limited to,

- **Presentation attack detection:** Deploying detectors at the ADC(s) to prevent presentation attacks at ①.
- **Continuous identity verification:**Continually verifying the actors identity during the session to ensure adversaries do not infiltrate the media capture process.
- **Device digital signature and integrity check implementation:** Enhancing the trustworthiness of the ADC(s), DAC(s), AI-powered system, and the target platform(s) by implementing digital signatures and integrity verification at ②, ⑥, ⑨, and ⑪. This will enable audiences and other stakeholders to authenticate the identities of devices and systems and prevent adversarial tampering.
- **Strong encryption:** Strongly encrypting the communication channels to prevent man-in-the-middle attacks at ③, ⑤, ⑦, ⑧, and ⑩.
- **Robust AI system protection:** Safeguarding AI-powered systems against various attacks such as adversarial attacks [2], backdoor attacks [12], and membership inference attacks [16] to counter potential threats at ⑥. These attacks can compromise system behavior and user data, enabling adversaries to create or inject deepfakes.
- **Input signal watermarking:** Watermarking signals captured by ADC(s) to prevent tampering during transmission to the AI-powered system, thereby addressing potential risks at ③. Traditional watermarking techniques [19] can be applied here.
- **Output media watermarking:** Watermarking media generated by AI-powered system [3] to enable tracking of manipulations and flagging of malicious alterations, addressing risks at ⑦, ⑧, ⑨, and ⑩.The watermarks can also be used to identify which AI-powered system was used to alter the content.

- **Additional information protection:** Encrypting or watermarking additional information in accordance with the context to prevent attacks at ④ and ⑤. Countermeasures are needed to verify the originality and integrity of this information.
- **Input media deepfake detection:** Implementing deepfake detection [18, 9] for media content input to the AI-powered system to ensure signals from the actor are authentic. If multiple AI-powered systems are used, the detected manipulations must align with the embedded watermarks from one system to the next.
- **Output media deepfake detection and watermark verification:** Implementing deepfake detection and watermark verification in the target platform(s) before their DAC(s) to ensure that final output does not contains deepfake manipulations, i.e., ones that go beyond legitimate alterations by the actor.
- **Purpose declaration and metadata embedding:** Requiring the actor to declare their intent to the AI-powered system, which embeds this information into the generated media content. Both the content and target platform(s) should clearly display the actor's intent and manipulation history. The target platform(s) should alert the audience(s) to malicious manipulations or inconsistencies in watermark and metadata information.

In summary, addressing the challenges posed by malicious manipulations, in the generative AI era requires a multi-faceted approach that goes beyond traditional detection methods. By integrating countermeasures at various points in the AI-powered framework, we can effectively safeguard against AI-generated attacks, ensure the integrity of media, and provide mechanisms for verifying the provenance, intention, and context of media content. These measures are essential for maintaining trust in AI-generated media content and protecting users from malicious manipulations in an increasingly complex digital landscape.

In the next two sections, we present two case studies that illustrate the two primary modes of generative AI usage: interactive and non-interactive content generation, each reflecting real-world applications. The interactive case study focuses on AI-powered communication systems, where real-time engagement between the actor and audience introduces unique security challenges. The non-interactive case study examines AI-driven content editing platforms where media manipulation occurs offline before distribution. For each scenario, we systematically select relevant attack vectors and corresponding defense strategies from our proposed countermeasure framework. By doing so, we provide a deeper, context-rich analysis that highlights the practical implications of these security measures and their effectiveness in mitigating AI-generated threats.

6 Case Study 1: AI-powered Communication Framework

In this case study, we provide an in-depth exploration of an interactive AI-powered communication framework inspired by advanced systems such as NVIDIA Maxine, as illustrated in Figure 6. This framework is derived from our proposed generalized AI-powered framework and is engineered to support high-fidelity, real-time two-way communication by capturing detailed voice signals and 3D video streams while dynamically tracking the spatial positions of participants. For simplicity, we depict only a single speaker and listener in a one-directional setting in Figure 6. Advanced computer vision algorithm facilitates continuous eye contact adjustment, and real-time audio translation with synchronized lip movements is employed to overcome language barriers. Additionally, the advanced enhancement algorithm improves video quality and compensates for missing frames and audio artifacts caused by low-bandwidth network conditions. These features are implemented via server-side processing using pre-trained AI models.

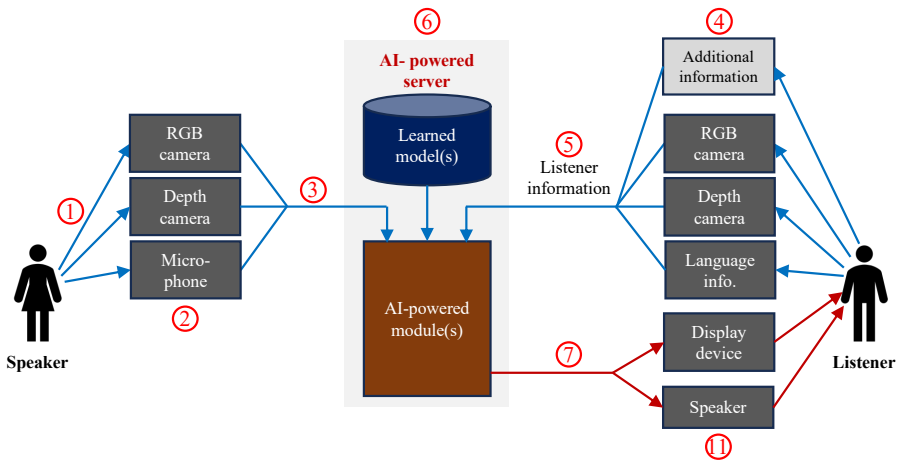


Figure 6: Potential critical points in an **interactive** AI-powered communication framework. Critical point numbers match those for generalized framework in Figure 5, with certain modules and critical points omitted to align with this specific design.

Beyond its core technical capabilities, the framework operates in a highly dynamic environment where multiple users engage in real-time interactions, often with shifting roles. This complexity introduces a wide range of security vulnerabilities that adversaries could exploit at different system levels. For instance, at the sensor level, malicious actors among the participants may manipulate video or audio feeds at the point of capture to perform presentation attacks. Adversaries may tamper with recording devices to manipulate data

or inject deepfakes to impersonate legitimate speakers. At the transmission level, adversaries may intercept communication channels to corrupt, alter, or eavesdrop on sensitive data exchanges. Additionally, the AI-powered system itself could become a point of compromise—either through external tampering, where adversaries modify its underlying models, training data, or workflows, or in more insidious cases, where the system itself behaves maliciously without the users' knowledge. These vulnerabilities highlight the critical need for a robust, multi-layered defense approach that goes beyond simple deepfake detection.

To counter these threats, the framework incorporates a suite of robust security mechanisms designed to safeguard the entire communication pipeline. Continuous identity verification ensures that each participant's legitimacy is maintained throughout the interaction, preventing unauthorized impersonation. Device digital signature checks authenticate the integrity of input sources, detecting any attempts to manipulate or replace original content. Sophisticated watermarking techniques provide an additional layer of security by embedding traceable markers into transmitted media, allowing post-analysis verification of authenticity. End-to-end encryption is implemented to secure communication channels, preventing unauthorized access or tampering during data transmission. To protect the AI-powered system from compromise, robust AI system protection measures, such as adversarial defenses, backdoor detection, and continuous integrity monitoring, are deployed. Similarly, the target platform undergoes strict digital signature and identity verification processes to ensure that the final content remains unaltered.

Table 1 summarizes the potential attacks and defenses at critical points within this communication framework. Notably, attacks and defenses at each point do not always have a direct one-to-one correspondence. Attackers may combine multiple attack strategies across different points, requiring the defender(s) to deploy a combination of countermeasures—some of which are interconnected—across multiple points to maximize protection. It is clear that detecting deepfake media is only one of many defense mechanisms and must be integrated with complementary strategies to comprehensively address the three crucial dimensions: provenance, intention, and context.

7 Case Study 2: Non-interactive AI-powered Framework for Content Editing

In this case study, we explore a non-interactive AI-powered framework designed for multimedia content editing—drawing inspiration from platforms such as Descript and Filmora—as depicted in Figure 7. Unlike interactive systems, where communication happens in real time, this framework enables an offline, one-way process where content creators modify multimedia con-

Table 1: Potential attacks and defenses at critical points in **interactive** AI-powered communication framework.

No.	Attack	Defense strategies
①	Presentation attack using deepfake material	+ Presentation attack detection + Continuous identity verification + Purpose declaration and metadata embedding (speaker)
②	Intercept ADCs to inject deepfake material	+ Device digital signature and identity verification + Input signal watermarking
③, ⑤, ⑦	Intercept channel to manipulate transmitted data	Strongly encrypt communication channels
④	Modify additional information	+ Non-media data encryption + Media data watermarking
⑥	Overwrite AI-powered system (model, training data, workflows, decision modules, etc.)	+ System digital signature and identity verification + Robust AI system protection + Input deepfake detection + Input watermark verification + Output media watermarking
⑪	Intercept ADCs & DACs	+ Device digital signature and identity verification + Deepfake and watermark verification + Metadata and auditing information display

tent before it is published. AI-powered tools assist in various editing tasks, such as adjusting transcripts, translating audio, and synchronizing visual elements, all of which are executed on high-performance server-side systems using pre-trained models. While these capabilities enhance efficiency and creative flexibility, they also introduce unique security challenges, as AI-assisted modifications can be exploited for malicious purposes, making it harder to differentiate between legitimate edits and deceptive manipulations.

A critical concern in this framework is the integrity of the media throughout its entire lifecycle, from acquisition to final distribution. Attackers can compromise source authenticity by manipulating or fabricating the initial media, thereby embedding malicious content before any AI processing occurs. During the editing phase, unauthorized modifications, metadata alterations,

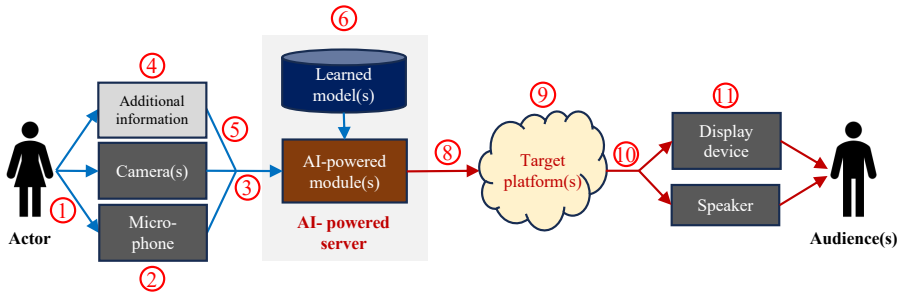


Figure 7: Potential critical points in a **non-interactive** AI-powered framework. Critical point numbers match those for generalized framework in Figure 5, with certain modules and critical points omitted to align with this specific design.

and malicious AI-generated content can be introduced—either by tampering with the AI-powered system itself or by exploiting weaknesses in the editing workflow. Once the final content is prepared for distribution, vulnerabilities in the target platform(s) may allow adversaries to alter or replace media content, misleading audiences and undermining trust. These risks highlight the necessity of a comprehensive security strategy that spans all stages of the content pipeline.

To mitigate these threats, the framework incorporates multiple layers of security measures. Source verification mechanisms ensure that only authentic media is used, reducing the risk of initial compromise. Digital signatures and cryptographic authentication techniques safeguard the integrity of edited content, preventing unauthorized modifications. Both input and output watermarking techniques are implemented to track content history and detect potential tampering. Metadata embedding allows for a transparent record of all modifications made during the editing process, creating an auditable trail that can be used to verify authenticity post-distribution. Additionally, rigorous AI system protection strategies–such as adversarial defenses, backdoor detection, and continuous monitoringare employed to prevent unauthorized manipulation of the AI-powered editing system itself. At the final stage, the target platform undergoes strict digital signature and identity verification to ensure that published content remains unaltered, thereby protecting audiences from exposure to malicious content.

Table 2 summarizes the potential attacks and defenses at critical points within this content editing framework. Similar to interactive systems, attacks and defenses do not always correspond in a simple one-to-one manner. Adversaries may combine multiple attack strategies across different stages of the content pipeline, requiring defenders to implement a combination of countermeasures—some of which are interdependent—to maximize security. Detecting deepfake media alone is insufficient; a holistic approach incorporat-

Table 2: Potential attacks and defenses at critical points in **non-interactive** AI-powered framework for content editing.

No.	Attack	Defense strategies
①	The actor is malicious	+ Identity verification + Purpose declaration and metadata embedding + Proactive defense (*e.g.*, watermarking) of published media content
②	Intercept ADCs to inject deepfake material	+ Device digital signature and identity verification + Input signal watermarking
③, ⑤, ⑧, ⑩	Intercept channel to manipulate transmitted data	Strongly encrypting communication channels
④	Modify additional information	+ Non-media data encryption + Media data watermarking
⑥	Overwrite AI-powered system (model, training data, workflows, decision modules, *etc*)	+ System digital signature and identity verification + Robust AI system protection + Source content inspection + Input deepfake detection + Input watermark verification + Output media watermarking
⑨	Overwrite target platform(s)	+ Platform digital signature and identity verification + Deepfake and watermark verification
⑪	Intercept DACs	+ Device digital signature and identity verification + Deepfake and watermark verification + Metadata and auditing information display

ing provenance verification, intention analysis, and contextual evaluation is essential.

8 Conclusions

We have shown that *traditional deepfake detection methods are insufficient in the generative AI era*, where both benign users and adversaries leverage advanced AI frameworks to create hybrid human-machine media content. The

significant advantages offered by generative AI coupled with its increasing acceptance within the community enable these "hybrid contents" to serve both positive and negative purposes depending on their context and presentation. To address these challenges, *we introduced a generalized framework for the application of generative AI, defined potential deepfake-related attacks, and presented systematically designed countermeasures that integrate various defense strategies.* We also presented two case studies demonstrating practical uses of the proposed countermeasures. Our aim is to harness the positive aspects of advanced AI technologies while minimizing their potential for misuse, particularly in the creation of malicious AI-generated material. Further research is needed to refine these countermeasures, explore additional attack vectors, and develop more sophisticated techniques to keep pace with the rapidly evolving landscape of generative AI.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grants JP-21H04907 and JP24H00732, by JST CREST Grants JPMJCR18A6 and JPMJCR20D3 including AIP challenge program, by JST AIP Acceleration Grant JPMJCR24U3, by JST K Program Grant JPMJKP24C2 Japan, by the project for the development and demonstration of countermeasures against disinformation and misinformation on the Internet with the Ministry of Internal Affairs and Communications of Japan, and by the "GENIAC (Generative AI Accelerator Challenge)" project, which is being implemented by the Ministry of Economy, Trade and Industry and NEDO with the aim of strengthening the development capabilities of generative AI in Japan.

References

[1] E. Altuncu, V. N. Franqueira, and S. Li, "Deepfake: definitions, performance metrics and standards, datasets, and a meta-review", *Frontiers in Big Data*, 7, 2024, 1400024.

[2] J. C. Costa, T. Roxo, H. Proença, and P. R. Inácio, "How deep learning sees the world: A survey on adversarial attacks & defenses", *IEEE Access*, 2024.

[3] S. Dathathri, A. See, S. Ghaisas, P.-S. Huang, R. McAdam, J. Welbl, V. Bachani, A. Kaskasoli, R. Stanforth, T. Matejovicova, *et al.*, "Scalable watermarking for identifying large language model outputs", *Nature*, 634(8035), 2024, 818–23.

[4] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, *et al.*, "The llama 3 herd of models", *arXiv preprint arXiv:2407.21783*, 2024.

[5] B. Egger, W. A. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, *et al.*, "3d morphable face modelspast, present, and future", *ACM Transactions on Graphics (ToG)*, 39(5), 2020, 1–38.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets", *Advances in neural information processing systems*, 27, 2014.

[7] S. Gupta, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Viable Threat on News Reading: Generating Biased News Using Natural Language Models", in *Fourth Workshop on Natural Language Processing and Computational Social Science - EMNLP*, 2020, 55–65.

[8] A. Hashmi, S. A. Shahzad, C.-W. Lin, Y. Tsao, and H.-M. Wang, "Understanding Audiovisual Deepfake Detection: Techniques, Challenges, Human Factors and Perceptual Insights", *arXiv preprint arXiv:2411.07650*, 2024.

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

[10] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models", *Advances in neural information processing systems*, 33, 2020, 6840–51.

[11] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A Watermark for Large Language Models", in *International Conference on Machine Learning*, PMLR, 2023.

[12] Y. Li, S. Zhang, W. Wang, and H. Song, "Backdoor attacks to deep learning models and countermeasures: A survey", *IEEE Open Journal of the Computer Society*, 4, 2023, 134–46.

[13] S. Marcel, J. Fierrez, and N. Evans, *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, Vol. 1, Springer, 2023.

[14] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature", in *International Conference on Machine Learning*, PMLR, 2023.

[15] A. Morales, G. Piella, and F. M. Sukno, "Survey on 3D face reconstruction from uncalibrated images", *Computer Science Review*, 40, 2021, 100400.

[16] J. Niu, P. Liu, X. Zhu, K. Shen, Y. Wang, H. Chi, Y. Shen, X. Jiang, J. Ma, and Y. Zhang, "A survey on membership inference attacks and defenses in Machine Learning", *Journal of Information and Intelligence*, 2024.

[17] N. K. Ratha, J. H. Connell, and R. M. Bolle, "An analysis of minutiae matching strength", in *Audio-and Video-Based Biometric Person Authentication: Third International Conference, AVBPA 2001 Halmstad, Sweden, June 6–8, 2001 Proceedings 3*, Springer, 2001, 223–8.

[18] C. Rathgeb, R. Tolosana, R. Vera-Rodriguez, and C. Busch, *Handbook of digital face manipulation and detection: From deepfakes to morphing attacks*, Springer Nature, 2022.

[19] M. Shaliyar and K. Mustafa, "Watermarking approach for source authentication of web content in online social media: a systematic literature review", *Multimedia Tools and Applications*, 83(18), 2024, 54027–79.

[20] J. Wu, R. Zhan, D. F. Wong, S. Yang, X. Yang, Y. Yuan, and L. S. Chao, "DetectRL: Benchmarking LLM-Generated Text Detection in Real-World Scenarios", in *Conference on Neural Information Processing Systems - Datasets and Benchmarks Track*.