IEEE Open Journal of the Computer Society

Received 12 July 2025; accepted 17 September 2025. Date of publication 1 October 2025;
date of current version 30 December 2025. The review of this article was arranged by Associate Editor Kai Liu.

Digital Object Identifier 10.1109/OJCS.2025.3616501

Generative AI-Driven Metaverse: The Promises and Challenges of AI-Generated Content

LAMIAA BASYONI [1], ADNAN QAYYUM [2], KHALED SHABAN [1], EZIEDDIN ELMAHJUB [3], ABDULAZIZ AL-ALI [1], OSAMA HALABI [1], AND JUNAID QADIR [1]

(Survey Paper)

[1]College of Computer Science and Engineering, Qatar University, Doha, Qatar
[2]Division of Information and Computing Technology, College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
[3]College of Law, Qatar University, Doha, Qatar

CORRESPONDING AUTHOR: ADNAN QAYYUM.

This work was supported in part by Qatar University High Impact Internal under Grant QUHI-CENG-23/24-127, in part by Qatar National Library, and in part by the Open Access funding is provided by the Qatar National Library.

ABSTRACT Recent advancements in Artificial Intelligence (AI) and immersive technologies, such as Extended Reality (XR), coupled with complementary innovations like 5G/6G wireless communications, are paving the way for fully realized AI-XR metaverses. AI will play a pivotal role in this transformation, enabling the seamless convergence of virtual and physical worlds. Among the various AI applications, Generative AI (GenAI) stands out for creating rich, dynamic, and interactive virtual environments—essential elements for sustained metaverse growth and user engagement. To facilitate a deeper understanding of how GenAI will be integrated and utilized within the metaverse, we provide a comprehensive overview of GenAI and Artificial Intelligence-Generated Content (AIGC). Specifically, we examine the potential influence of GenAI's on future metaverses, exploring both the opportunities it offers and the major challenges associated with its deployment. Additionally, we investigate the robustness of AIGC detection techniques against adversarial attacks, highlighting less-explored risks posed by adversarial examples. Finally, we highlight various open research issues related to GenAI, AIGC, the metaverse, and responsible innovation that merit further exploration.

INDEX TERMS Artificial intelligence, extended reality, foundation models, generative AI, generative pre-trainer transformer, immersive experience, machine learning, virtual reality.

I. INTRODUCTION

In recent years, Generative Artificial Intelligence (GenAI) has revolutionized content creation, leading to the emergence of Artificial Intelligence-Generated Content (AIGC), which encompasses a vast array of multi-modal content. GenAI techniques have garnered substantial attention from both the computer science community andvarious industries, as demonstrated by popular systems such as ChatGPT,[1] Bard,[2] and DALL-E,[3] among others. GenAI plays a vital role in facilitating efficient content creation for the metaverse. By emulating human cognitive abilities, such systems can generate a wide range of AIGC, including intelligent Non-Player Characters (NPCs), automated question-answering, dialogues, and even digital humans (e.g., digital twins and avatars). The integration of GenAI technologies will enhance the immersive experience of the metaverse by offering interactive and intelligent entities with which users can engage.

AIGC solutions revolutionize content creation by automating large-scalre generation across various domains. This capability is particularly beneficial for industries that require frequent updates or have high demands for content production, such as news agencies, social media platforms, and e-commerce websites. Moreover, these solutions optimize operations by saving time and reducing effort, ensuring a

[1]https://chatgpt.pro/
[2]https://bard.google.com/chat
[3]https://openai.com/dall-e-2

TABLE 1. Comparison of Our Paper With Existing Survey and Review Articles. Legend: S: Security; P: Privacy; ✓: Covered; ✗: Not Covered; ≈: Partially Covered.

The Work	Year	Contribution(s)	Metaverse Basics	Metaverse S&P	AIGC Fundamentals	AIGC in the Metaverse	AIGC Detection
Sun et al. [1]	2022	Introduces the basic concepts, definition, technology, and privacy issues of the metaverse.	✓	✓	✗	✗	✗
Wang et al. [2]	2022	Survey of the fundamentals, security, and privacy of the metaverse.	✓	✓	✗	✗	✗
Anantrasirichai et al. [3]	2022	Categorizes the creative applications based on incorporating AI technologies.	✗	✗	✓	✗	✗
Ning et al. [4]	2023	Explores policies of various countries, companies, and organizations relevant to the metaverse.	✓	✓	✗	✗	✗
Zhang et al. [5]	2023	Reviews ChatGPT with its underlying technology, applications, and challenges.	✗	✗	✓	✗	✗
Lee et al. [6]	2023	Discusses from a user-centric perspective how the metaverse development will accelerate with AIGCs.	✓	✗	✓	✓	✗
Wu et al. [7]	2023	Provides an outlook on the future development of deepfake in the metaverse.	✓	✗	✓	✓	✗
Tariq et al. [8]	2023	Discusses how deepfakes can be used to impersonate in specific scenarios in the metaverse.	✓	✗	✓	✓	✗
Wu et al. [9]	2023	Overview of AIGC and examine AIGC's potential integration with the metaverse.	✓	✗	✓	✓	✗
Basyoni et al. [10]	2023	Provides a brief overview of AIGC risks and detection techniques in the metaverse context.	✗	≈	✗	≈	≈
Wang et al. [11]	2024	Discusses AIGC's role in metaverses including a brief discussion on challenges and mitigation techniques.	✗	✗	✓	✓	≈
Xu et al. [12]	2023	Explores caching and inference frameworks for AIGC services in edge-based metaverse applications	✓	≈	✓	✓	✗
Chen et al. [13]	2024	Explores AIGC-driven digital twin networks to enhance metaverse immersion and intelligence.	✓	≈	✓	✓	✗
This Work	-	Reviews state-of-the-art AIGC technologies, their role and risks in the metaverse, and the relavant detection methods.	✓	✓	✓	✓	✓

consistent and dynamic flow of fresh and engaging material. Among the various methods of AIGC, text-guided content generation stands out as particularly practical, enabling meaningful interaction between human instructions and the AI system for precise and tailored content output.

However, the rising popularity of GenAI has also led to its exploitation for malicious purposes, highlighting the pressing need to develop effective methods for detecting and identifying misleading content. Consequently, detecting AIGC has become a crucial research focus, as it underpins the integrity, authenticity, and trustworthiness of content across diverse contexts. Addressing this challenge requires continuous research and innovation to stay ahead of emerging threats and safeguard against potential risks associated with AIGC misuse. In this paper, we provide a comprehensive survey of GenAI and AIGC in metaverses. Specifically, we explore the potential benefits of AIGC for enhancing user experiences within the metaverse while also examining scenarios where such content can pose risks or cause harm. Furthermore, we review various techniques employed to detect different types of AIGC, thereby supporting the identification and mitigation of potential concerns arising from its usage.

A. RELATED STUDIES

The metaverse has emerged as a transformative domain, capturing the attention of researchers and practitioners across diverse industries. Numerous studies have explored its foundational concepts, technological underpinnings, and the security and privacy challenges it entails [1], [2], [4], [14]. Complementary to these efforts, a growing body of research has examined AIGC, focusing on its potential applications, enabling technologies, and detection mechanisms [5], [6]. However, as highlighted in Table 1, there remains a critical gap in the literature—a comprehensive survey that synthesizes the implications of AIGC in the metaverse context.

While prior works provide valuable insights into specific aspects, including AIGC applications [9], creative uses of AI technologies [3], and the risks posed by deepfakes in metaverse scenarios [7], [8], they have not yet presented an integrated view of AIGC's role, risks, and detection strategies within the metaverse. Recent studies, such as [11], emphasize the potential of AIGC for creating immersive metaverse experiences but do not address detection techniques or the broader challenges of integrating AIGC safely and ethically into virtual environments.

In this context, our work bridges this gap by synthesizing existing literature, offering a comprehensive taxonomy, and identifying key challenges and promising directions for future research. We critically evaluate the current state of AIGC technologies and their applications within the metaverse, while emphasizing the importance of detection techniques and mitigation strategies to address emerging risks.

B. CONTRIBUTIONS AND ORGANIZATION OF THIS PAPER

Our study not only builds on existing works, such as those by Xu et al. [12] and Chen et al. [13], which focus on specific technical innovations, but also integrates their insights into a broader framework. By doing so, we provide a consolidated foundation for understanding the interplay between AIGC and the metaverse, offering valuable guidance for future research and development in this domain. The key contributions of our study can be summarized as follows:

1) *Comprehensive Review of GenAI in the Metaverse:* We provide an in-depth review of GenAI technologies and their applications within the metaverse. This includes an exploration of capabilities, recent advancements, emerging trends, and the transformative potential of GenAI in shaping the metaverse ecosystem.
2) *Integration of AIGC in the Metaverse:* We examine the integration of AIGC into the metaverse, highlighting its opportunities and challenges. Our analysis covers the benefits of leveraging AIGC for immersive experiences,

TABLE 2. List of Acronyms

Acronym	Definition
5G/6G	Fifth/Sixth Generation Wireless Communication
AI	Artificial Intelligence
AIGC	Artificial Intelligence-Generated Content
AR	Augmented Reality
CCPA	California Consumer Privacy Act
CNN	Convolutional Neural Network
DDPM	Denoising Diffusion Probabilistic Model
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
GPT	Generative Pretrained Transformer
LLM	Large Language Model
LSTM	Long Short-Term Memory
MCMC	Markov Chain Monte Carlo
NCE	Noise Contrastive Estimation
NLP	Natural Language Processing
PFM	Pretrained Foundation Model
RLHF	Reinforcement Learning from Human Feedback
RNN	Recurrent Neural Network
VAE	Variational Autoencoder
VR	Virtual Reality
XR	Extended Reality

while addressing risks related to security, privacy, and ethical concerns.

3) *Evaluation of AIGC Detection Techniques:* We critically assess state-of-the-art methods for detecting AIGC in the metaverse. By discussing the strengths and limitations of existing approaches, we provide insights into their efficacy and practical implications for ensuring safety and integrity in Metaverse environments.
4) *Robustness and Gaps in AIGC Detection:* We investigate the resilience of AIGC detection methods against evolving threats and adversarial scenarios. Our work identifies gaps and opportunities for improvement, offering actionable recommendations for enhancing the reliability and accuracy of AIGC detectors.
5) *Guidelines for Responsible Innovation:* Building on our findings, we propose actionable guidelines for the responsible adoption of GenAI and AIGC in the metaverse. These guidelines aim to balance innovation with safeguards, addressing technical, ethical, and societal considerations to foster a safe and inclusive metaverse.

Organization of the Paper: The remainder of the paper is organized as follows. Section II provides background on the underlying technologies driving AIGC. Section III introduces the fundamental aspects of the metaverse and explains the role AIGC within this virtual environment. Section IV discusses AIGC detection methods in the metaverse and reviews the robustness of these detection techniques. Section V outlines various open research issues. Finally, Section VI concludes the paper with a summary. For quick reference, Table 2 provides a comprehensive list of acronyms used throughout this paper.

II. BACKGROUND

A. THE METAVERSE: AN INTRODUCTION

The term *metaverse* was originally introduced by science fiction writer Neal Stephenson in his 1992 novel *Snow Crash*. In this work, Stephenson envisioned the metaverse as a three-dimensional digital environment—essentially a shared virtual reality (VR) experience where users could seek refuge from a mundane physical world. The term "metaverse" is derived from the prefix "meta," signifying transcendence, and the suffix "verse," representing a universe. It describes a digitally generated environment with its own cohesive set of values and an autonomous economic system interconnected with the physical world. In a recent book, Graylin et al. [15] defined the metaverse as a 3D global internet powered by AI and accessed through XR devices. Mathematically, this can be expressed as:

$$\text{Metaverse} = \text{Internet}(3\text{D})^{\text{AI}} \leftrightarrow \text{XR}$$

In recent years, the metaverse has evolved into a vast virtual ecosystem, garnering the interests of technology enthusiasts, entrepreneurs, and researchers alike. Motivated by practical needs and the potential feasibility of building metaverses at scale, global interest in the concept has surged. Several major technology companies—including Facebook (rebranded as "Meta" [16]), Microsoft, Tencent, and NVIDIA—have declared their active involvement in metaverse development initiatives.

Despite its growing popularity, the metaverse lacks a universally accepted definition. It is often broadly described as the convergence of physical, augmented, and virtual realities, enabling immersive interactions between real and simulated environments [4], [17]. Some perspectives portray the metaverse as a fully immersive, hyperspatiotemporal, and self-sustaining shared virtual space that merges physical, human, and digital dimensions. Moreover, it is frequently positioned as an evolution of the Internet, allowing users to participate in persistent digital environments through embodied experiences and alternate identities [2]. In this paper, we define the metaverse as a digitally mediated environment that enables continuous, immersive interactions by integrating real and virtual environments, supported by enabling technologies such as extended reality (XR), blockchain, and Generative AI (GenAI). In the metaverses, GenAI will play a crucial role in generating dynamic content, personalizing user experiences, and facilitating real-time interactions, thereby enhancing the richness, adaptability, and scalability of metaverse applications.

Looking ahead, the metaverse is expected to spawn numerous promising applications, including educational metaverses [18], surgical metaverses [19], and industrial metaverses [20], and as well as use cases in autonomous vehicles [21], [22] and traffic safety [23]. A bibliometric analysis of 928 articles outlining the evolving research landscape of the metaverse is presented in [24].

B. AN INTRODUCTION TO GENAI

GenAI represents a transformative paradigm in artificial intelligence, capable of producing diverse and realistic content—such as text, images, and videos—based on user-provided prompts. GenAI models often comprise a generator, which creates novel outputs, and a discriminator, which evaluates

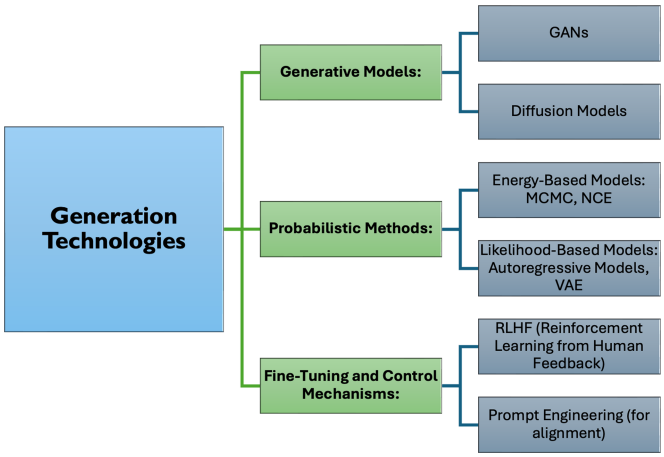

FIGURE 1. **GenAI Generative Models and Techniques.** ***Highlights the foundational generative approaches of GenAI, from GANs for high-fidelity content to diffusion models and fine-tuning mechanisms like RLHF, tailored for scalable and diverse applications.***

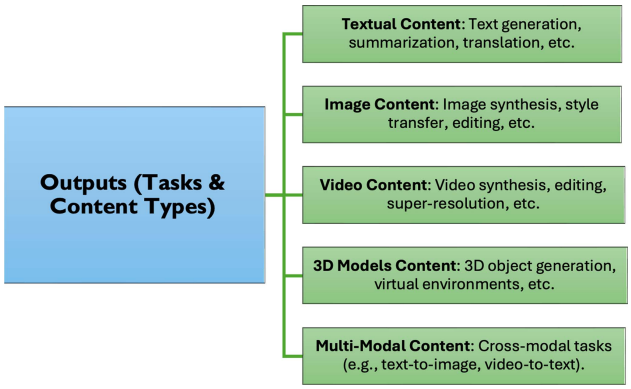

FIGURE 2. **GenAI Output Types and Tasks.** ***These capabilities highlight GenAI's transformative potential across industries, enabling innovative solutions in content creation, virtual environments, and interactive experiences.***

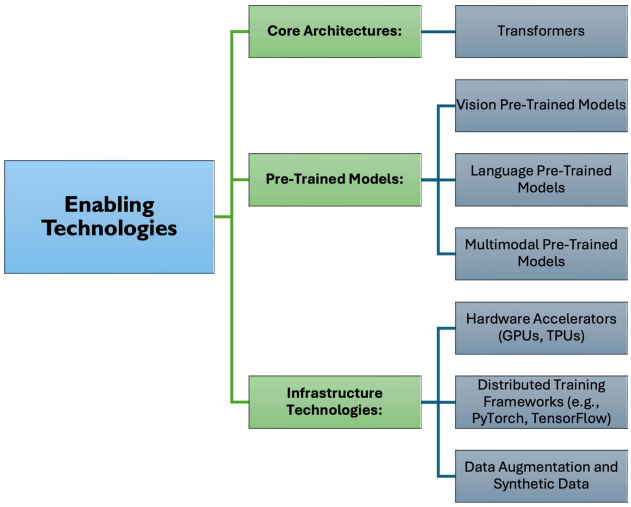

FIGURE 3. **GenAI Enabling Technologies.** ***Depicts the foundational architectures, models, and hardware that underpin the scalability and performance of GenAI systems.***

their authenticity. This setup, combined with stochastic behavior, ensures that even when identical prompts are used, the generated content may vary.

Modern GenAI is grounded in the Transformer architecture introduced by Vaswani et al. in 2017 [25], which revolutionized Natural Language Processing (NLP) through its attention mechanism. This mechanism enables the model to selectively focus on crucial segments within input sequences. Over time, Transformers have been adapted to multimodal tasks, extending their reach to domains such as Computer Vision [26], and have become a foundational technology for generative models across various fields.

The evolution of GenAI has been punctuated by three key inflection points, as noted by Horvitz and Mitchell [27]:

- The rise of deep learning, serving as a catalyst for rapid innovation
- The integration of transformer-based self-supervised learning on vast, web-scale datasets
- The development of reinforcement learning from human feedback (RLHF), enabling models to follow natural language instructions and sustain contextually relevant conversations

AIGC is the application of GenAI to generate content tailored to user-defined instructions. AIGC systems rely on three fundamental pillars: algorithms, data, and computing resources. High-quality multimodal datasets (e.g., text, images, and audio) are essential for model training, while cutting-edge hardware (e.g., advanced GPUs) and cloud infrastructures facilitate the operation of large-scale architectures with billions of parameters [9]. These integrated components enable the seamless generation of personalized, contextually relevant content.

Figs. 1, 2, and 3 present a detailed taxonomy of GenAI for AIGC, illustrating the interply among the enabling technologies, generative techniques, and content creation tasks. These figures provide a structured framework for understanding the core components and hierarchical organization of GenAI, setting the stage for its diverse applications. The following sections reviews these elements, categorizing various models according to their underlying mechanisms and the tasks they address.

C. TYPES OF GENAI MODELS AND GENERATION TECHNOLOGIES

Fig. 1 presents a taxonomy of GenAI generation technologies, classifying them into three broad categories: generative models, probabilistic methods, and fine-tuning mechanisms. Each category encompasses various techniques and architectures that enable GenAI systems to produce high-quality outputs.

1) GENERATIVE MODELS

Generative models learn the underlying data distribution and directly produce new data samples. Prominent examples include:

- *Generative Adversarial Networks (GANs):* GANs consist of a generator-discriminator pair, where the generator creates synthetic samples and the discriminator evaluates their realism. Advances such as StyleGAN and BigGAN have considerably improved output quality, particularly in text-to-image synthesis [28].
- *Diffusion Models (DMs):* Diffusion models such as DDPM [29], employ a noise-injection and denoising process to produce realistic outputs. Renowned for their stability and versatility, they are widely applied in image and video generation.

2) PROBABILISTIC METHODS

Probabilistic methods optimize objectives based on likelihood or energy functions to model data distributions:

- *Likelihood-Based Models:* These include autoregressive approaches like GPT and flow-based methods like Glow. Variational Autoencoders (VAEs) also fit in this category, learning smooth latent-space representations.
- *Energy-Based Models (EBMs):* EBMs define data probabilities using energy functions. Training can involve techniques such as Markov Chain Monte Carlo (MCMC) and Noise Contrastive Estimation (NCE), although these can be computationally demanding.

3) FINE-TUNING AND CONTROL MECHANISMS FOR IMPROVING AIGC QUALITY

Fine-tuning and control mechanisms are essential for customizing GenAI models to specific tasks and enhancing AIGC quality. They address challenges like aligning model outputs with user intent, improving content utility, and ensuring factual correctness.

a) Reinforcement Learning with Human Feedback (RLHF): RLHF has emerged as a powerful approach for aligning model outputs with human preferences, especially in large language models like ChatGPT [30]. First introduced in [31], RLHF involves training a reward function based on human feedback and optimizing the model to maximize this reward. By incorporating human guidance, RLHF allows models to learn optimal policies and refine their outputs, making them more useful and aligned with user expectations.

b) Prompt Engineering: Prompt engineering involves carefully crafting input prompts to guide model behavior and elicit the desired outputs. This technique is integral to maximizing GenAI's practical usability, ensuring that generated content aligns with user requirements while minimizing additional fine-tuning.

By combining these fine-tuning strategies with generative and probablistic methods, GenAI models can be effectively tailored to specific domains or contexts, improving the overall quality of AIGC across diverse applications.

D. TYPES OF GENAI TASKS

GenAI tasks cover a broad spectrum of applications and outputs, categorized according to modality and the nature of the generated content. Fig. 2 summarizes these tasks, dividing them into textual, image, video, 3D models, and multimodal content categories.

- *Textual Content:* Tasks include text generation, summarization, translation, and creative writing.
- *Image Content:* Tasks span image synthesis, style transfer, and editing.
- *Video Content:* Use cases include video synthesis, super-resolution, and animation generation.
- *3D Models Content:* Focuses on generating virtual objects and environments, as well as reconstructing 3D models.
- *Multi-Modal Content:* Integrates data from multiple modalities, such as text-to-image, image-to-text, and video-to-text.

Below, we briefly discuss unimodal and multimodal tasks.

1) UNIMODAL TASKS

Unimodal tasks involve generating outputs within the same modality as the input. They form foundational of many GenAI systems and include:

- *Text Generation:* Large Language Models (LLMs) like GPT [32] and T5 [33] perform summarization, machine translation, and question answering. These models generate coherent, contextually relevant text, enabling applications in content creation, knowledge extraction, and conversational AI.
- *Image Generation:* Unimodal image tasks include restoration, enhancement, and style transfer. Models like StyleGAN and diffusion-based methods produce high-quality visuals, supporting applications such as semantic editing, artistic rendering, and restoring degraded images.
- *Video Content:* Tasks such as video synthesis and animation fall under this category. Advanced techniques aim to maintain temporal coherence and smooth transitions between frames, enabling realistic animations and enhanced video quality through super-resolution methods.
- *3D Models Content:* 3D generation tasks create virtual objects and environments, crucial for augmented reality, gaming, and metaverse development. For example, Neural Radiance Fields (NeRF) [34] enable realistic 3D model generation from input data.

2) MULTIMODAL TASKS

Multimodal tasks integrate and translate information across different modalities, making them essential for more complex GenAI applications. Examples include:

- *Text-to-Image:* Models like DALL-E [26] generate high-quality images from textual descriptions, leveraging autoregressive or diffusion-based approaches to align text and visual semantics.
- *Image-to-Text (Image Captioning):* Methods such as Neural Image Caption (NIC) [35] and VisualBERT [36] combine image and text representations to produce

descriptive captions, enhancing tasks like visual storytelling and accessibility.
- *Video-to-Text:* Generating textual summaries or annotations for video requires extracting semantics across multiple frames, aiding in video indexing and summarization.
- *Text-to-3D:* Advanced systems like DreamField [37] and CLIP-Mesh [38] extend text-to-image principles to produce detailed 3D models and immersive environments, supporting metaverse and industrial design.
- *Cross-Modal Analysis:* These tasks evaluate and synthesize relationships among modalities, such as combining video and audio or mapping textual descriptions onto visual content, enhancing the interpretability and versatility of GenAI systems.

E. GENAI ENABLING TECHNOLOGIES

Fig. 3 highlights the foundational enabling technologies that power GenAI systems. These include core architectures, pre-trained models, and infrastructure solutions, collectively supporting scalable, efficient, and high-performance generative capabilities. The following subsection provide further detail on each of these components.

1) ADVANCEMENTS IN TRADITIONAL ML/DL MODELS

The transformer architecture has become the dominant backbone in modern NLP, largely replacing recurrent neural networks (RNNs). For computer vision, the Vision Transformer (ViT) and traditional convolutional neural networks (CNNs) both continue to thrive [39]. Below, we briefly discuss these fundamental architectures and their roles in the evolution of GenAI.

a) RNN Architecture: RNNs were foundational in NLP tasks by capturing sequential dependencies in temporal data such as language or audio. As RNNs feature cyclic connections, allowing previous outputs to serve as inputs and facilitating memory retention. Different RNN configurations—one-to-one, many-to-one, one-to-many, and many-to-many—address a variety of tasks [40]. However, RNNs struggle to model long-range dependencies due to gradient propagation issues. To alleviate this, Long Short-Term Memory networks (LSTMs) were introduced, enabling selective retaintion and disposal of information through mechanisms known as gates [41].

b) CNN Architecture: CNNs are specialize in processing grid-like data, such as images, and draw inspiration from visual cortex. They consist of convolution, pooling, and fully connected layers. Convolution layers perform feature extraction through specialized linear operations, pooling layers reduce dimensionality, and fully connected layers map learned features to predictions [42]. Pioneering architectures like AlexNet demonstrated the impact of depth and large datasets [43], followed by subsequent innovations such as VGGNet, ResNet, and GoogLeNet (Inception), which advanced performance while reducing resources demands [44], [45].

c) Generative Adversarial Networks (GANs): Introduced by Goodfellow et al. [46], GANs are central to generative AI. They consist of a *generator* that creates synthetic data and a *discriminator* that distinguishes between real and generated samples. This adversarial training framework powers numerous applications, including image synthesis, text-to-image generation, and style transfer. Improved variants like StyleGAN and BigGAN have have achieved photorealistic images quality. However, GANs also face challenges, such as mode collapse and computational intensity during training. Ongoing research explores solutions to enhance stability, output diversity, and security—critical aspects for metaverse applications [28]. In the metaverse, GANs facilitate tasks such as generating realistic avatars, dynamic virtual environments, and complex textures, which are pivotal for creating immersive and adaptive virtual spaces. Moreover, GANs are being explored for security-critical applications in the metaverse, such as protecting user data from black-box model inversion attacks, ensuring both creativity and security in virtual ecosystems [28].

2) TRANSFORMERS AND PRETRAINED FOUNDATION MODELS

a) Transformers: The transformers have revolutionized deep learning by relying solely on attention mechanisms, eliminating the need for recurrence or convolutions [25]. Their abiltiy to focus on relevant parts of a sequence, irrespective of positional distance, enables parallelization. In the transformer architecture, both the encoder and decoder are composed of stacked self-attention and point-wise, fully connected layers. While the encoder processes input representations, the decoder generates outputs by additionally referencing encoder-produced context vectors.

b) Pretrained Foundation Models (PFMs): In the big-data era, PFMs have become indispensable in AI. by training on large, often unlabeled datasets, PFMs learn learn rich feature representations applicable across varied tasks, such as text classification, generation, image classification, object detection, and graph analysis [47]. In NLP, BERT citedevlin2018bert and GPT [32] exemplify two major fine-tuning paradigms. BERT employs a bidirectional encoder from both preceding and subsequent tokens, which is effective for comprehension tasks but less suited for geneerative tasks. GPT, on the other hand, uses an autoregressive decoder to predict future tokens based on past input, making it well-suited for text generation. A notable instance of a PFM is ChatGPT, which leverages RLHF to align the GPT-3.5 model's outputs with human preferences, enhancing both utility and coherence.

F. GENAI IN METAVERSES

Ongoing advancements in immersive technologies, such as VR and AR, have driven rising interest in metaverse development. Despite substantial investments, however, the

scale of metaverse content remains insufficient to meet user demand, largely due to high production costs and technical barriers. As a result, even well-funded projects often lack the diversity, openness, and refinement needed for truly compelling experiences.

1) REDUCING BARRIERS TO METAVERSE DEVELOPMENT

GenAI offers a transformative solution for automating and streamlining content creation within the metaverse. The ability to convert simple textual prompts into complex digital assets—such as 3D models, virtual environments, and interactive avatars, significantly lowers the technical barriers traditionally associated with metaverse development. This democratization of content creation allows individuals and small development teams, even without advanced programming or design expertise, to contribute meaningfully to immersive digital environments. Practical applications of this capability are already evident in tools such as OpenAI's Point-E for 3D object generation [48] and NVIDIA's Omniverse ACE for AI-driven character animation [49]. These systems enable faster creation of digital content and reduce the costs associated with traditional development pipelines. Moreover, user-generated content platforms including Roblox[4] and Fortnite Creative[5] have demonstrated how non-expert users can design interactive worlds and experiences at scale, highlighting a measurable reduction in both time and technical barriers.

2) INTEGRATING GENAI WITH ENABLING TECHNOLOGIES

The impact of GenAI extends beyond content generation, intersecting with several key enablers of the metaverse:

- *Digital Twin Technology:* GenAI leverages digital twins to construct virtual counterparts of real-world environments, enhancing realism and immersion [13].
- *Immersive Technologies:* VR and AR deliver rich 3D experiences, while GenAI ensures that content within these spaces is dynamic, personalized, and responsive.
- *Next-Generation Wireless Networks:* Low-latency, high-reliability connections offered by 5G/6G support seamless metaverse experiences and real-time GenAI-powered interactions [50].

By combining these technologies, GenAI becomes the backbone for adaptive, high-fidelity metaverse applications.

3) PERSONALIZATION AND DYNAMIC ADAPTABILITY

GenAI significantly enhances immersion and interactivity in the metaverse:

- *Personalized Avatars:* Automatically generated avatars can reflect a user's unique characteristics and preferences.
- *Dynamic Virtual Environments:* GenAI can adjust virtual spaces in real time based on user interactions, behavioral cues, and sensor data (e.g., from brain-computer interfaces).

[4]https://create.roblox.com/landing

[5]https://www.fortnite.com/create?lang=en-US

FIGURE 4. The role of GenAI in facilitating diverse AIGC use cases within metaverses.

- *Continuous Engagement:* Such adaptability ensures that virtual spaces remain relevant, captivating, and ever-evolving, sustaining user interest over time.

4) THE ROLE OF GENAI IN METAVERSE EXPANSION

GenAI stands poised to be a primary driving force behind the metaverse's ongoing evolution. By addressing core content-creation challenges, enabling dynamic adaptability, and integrating with complementary technologies, GenAI unlocks experiences once considered unfeasible. These capabilities will not only spur innovation but also ensure the long-term growth and viability of metaverse platforms.

The following section (Section III) takes a closer look at how GenAI is applied within the metaverse, providing deeper insights into how these models and techniques enhance immersive environments, facilitate content creation, and tackle challenges associated in large-scale virtual ecosystems.

III. THE ROLE OF GENAI IN THE METAVERSE

Recent innovations in GenAI have significantly expanded the frontiers of content creation, offering the potential to transform the metaverse in various dimensions (see Fig. 4). GenAI is now capable of generating a wide range of high-quality content—from images and videos to entire virtual worlds—dramatically enhancing realism and immersion. The integration of cutting-edge GenAI technologies from both industry and academia holds the promise of unlocking the full potential of AIGC, fostering creativity, and broadening the scope of metaverse applications. By enabling the dynamic, real-time generation of virtual environments that respond to user input, GenAI supports personalized and adaptive experiences in the metaverse. Moreover, the deployment of intelligent non-player characters (NPCs) equipped with natural interaction capabilities further amplifies immersion. Beyond and interactivity. Beyond gaming, GenAI-driven applications also extend to industries like architecture, product design, education, and training simulations, offering a diverse array of use cases [6], [51], [52], [53].

This section discusses GenAI's transformative role in the metaverse, focusing on capabilities in generating virtual worlds, avatars, and interactive NPCs, alongside broader applications in domains such as education and training. It also

addresses the challenges associated with integrating GenAI into metaverse ecosystems, including ethical considerations and technical limitations. Collectively, these insights highlight both the opportunities and complexities of leveraging GenAI for immersive, sustainable virtual worlds.

A. GENAI FOR METAVERSE CONTENT CREATION

AIGC is pivotal in efficiently generating content within the metaverse, spanning intelligent NPCs, chatbots, dialogue systems, and digital twins. This breadth of content enriches user engagement by offering interactive and intelligent entities [9].

1) VIRTUAL WORLD AND CONTENT GENERATION

Creating robust virtual worlds is resource-intensive, often requiring substantial development efforts even for short-lived experiences, particularly when user-generated content is needed. To address these challenge, researchers increasingly adopt GenAI techniques that democratize and streamline virtual content production [54]. Recent advancements aim to generate diverse and immersive environments suited to various contexts. For example, Chen et al. [55] introduced a text-driven framework for generating realistic, high-dynamic-range panoramas, facilitating immersive VR scenes from textual descriptions. In another instance, Huang et al. [56] proposed assigning realistic audio to static panorama images by combining scene classification, object depth estimation, and audio-sources placement. Similarly, Prpa et al. [57] combined immersive virtual environments with respiration sensors and generative soundscapes to produce meditation experiences synchronized with user breathing patterns.

By integrating GenAI models into the metaverse platforms, developers can innovate in several dimensions, improving both content creation and user experiences:

- *3D Modeling and Virtual Object Creation:* GenAI facilitates users-friendly virutal object design (e.g., buildings, furniture), significantly speeding up content generation.
- *Enhanced 3D Object Detection:* GenAI-based algorithms bolster real-time, context-aware perception of virtual elements, enhancing augmented reality applications.
- *Artistic Design and Customization:* These systems empower creators to craft intricate, personalized objects, encouraging creativity in virtual spaces.
- *2D-to-3D Reconstruction:* GenAI tools transform 2D images into interactive 3D models, enriching cultural and educational experiences like virtual museums.
- *Point Cloud Generation and Shape Modeling:* For professional use cases in architecture and design, GenAI accelerates point-cloud generation and and virtual architectural modeling, enabling rapid exploration of new designs.

Through such applications, GenAI enriches the metaverse with more realistic, diverse, and interactive 3D content [58], enhancing user engagement while expanding possibilities for creative expression.

2) AVATARS AND NON-PLAYER CHARACTERS (NPCS)

Avatars and NPCs are central to user immersion and engagement within the metaverse. Below, we explore the advancements, use cases, and potential pitfalls in integrating GenAI for these virtual identities.

a) Realistic Avatar and NPC Creation: Realistic avatars and NPCs are integral to constructing lifelike, engaging metaverse environments. GenAI has made significant contributions in this domain:

- *Avatar Generation:* Nagano et al. [59] used GANs for lifelike avatar creation, while Wang et al. [60] designed a generative model to synthesize highly realistic human head images with consistent multi-view renderings.
- *NPC Design:* Efforts include rendering, facial expression generation, motion synthesis, and behavior modeling [61], [62], [63], demonstrating how GenAI can create dynamic, responsive NPCs.

b) Enhancing NPC Behavioral Traits: GenAI endows NPCs with greater adaptability and diverse behavioral profiles:

- Dobre et al. [64] demonstrated how machine learning (ML) enables NPCs to respond dynamically to in-game events, transcending scripted behaviors.
- Personalized interactions become feasible as NPCs adapt to player actions and dialogue choices, resulting in richer, more personalized user experiences [65], [66].

c) Emerging Interaction Technologies: Technologies such as brain-computer interfaces promise to redefine how users interact with NPCs and virtual worlds:

- These interfaces may permit thought-based research envisions "invisible" human-computer interfaces, rendering interactions fluid and transparent [67], [68], [69].

d) Applications in Gaming: GenAI functions as an intelligent agent in gaming, transforming the player experience:

- *Customization:* Players can tailor game scripts, character appearances, and emotional expressions, yielding highly personalized gameplay [70].
- *Dynamic Content Creation:* Developers can employ GenAI to simulate player behavior, infuse new content, and adjust game difficulty levels.

e) Ethical and Security Concerns: While GenAI introduces substantial benefits, it also poses ethical and security challenges. Overreliance on AI-driven processes may foster malicious or unethical activities if not properly regulated [71]. Integrating ethical guidlines into GenAI applications is critical to maintaining a safe, equitable metaverse ecosystem.

B. GENAI FOR OTHER METAVERSE DOMAINS

Beyond content creation for avatars and NPCs, GenAI holds promise for additional metaverse sectors, including automatic digital twins and personalized experiences. However, dedicated research on these specialized use cases remains limited [51]. In metaverse authoring, AIGC can enable creative individuals to generate partial or fully realized digital objects, boosting efficiency and innovation. Much of the current research has centered on visual content creation, such as 3D

reconstruction, creative assistance, 3D molecule modeling, and free-viewpoint video generation [72], [73].

Within the fast-evolving GenAI landscape, pioneering systems are merging AIGC with user-generated content to enhance interactivity. For example, Wander is a prototype AI chatbot that uses AIGC to create interactive storytelling experiences [74]. The system integrates AIGC with user-submitted text via social messaging platforms, leveraging knowledge graphs to orchestrate dynamic narratives. Though AIGC is is just one aspect of AI in the metaverse, AI more broadly is expected to underpi a wide spectrum of applications across various layers of metaverse infrastructure [15].

C. CHALLENGES ASSOCIATED WITH GENAI IN THE METAVERSE

While GenAI presents significant opportunities for creating immersive and dynamic virtual environments, its integration into the metaverse introduces a range of legal, technical, and ethical challenges that require careful management. This section outlines the major areas of concern and highlights potential directions for addressing them.

1) JURISDICTION AND LAW ENFORCEMENT

The global, borderless nature of the Metaverse complicates traditional legal frameworks, which typically assume geographically defined jurisdictions [75]. Ownership and governance of these virtual environments is often controlled by multinational corporations, raising critical questions about which legal systems apply in cases of dispute or misconduct. Moreover, enforcing civil or criminal liability is further complicated by the anonymity of user avatars and possibility of individuals operating multiple virtual identities. In the relevant literature, a few researchers such as Qin et al. [76] have advocated for the development of an international governance framework capable of harmonizing cross-border legal efforts to regulate metaverse platforms. However, the practical implementation of such systems in decentralized environments remains unresolved. One potential solution could be to adopt a jurisdiction-by-design approach, where legal compliance mechanisms are embedded within metaverse platforms based on the users' physical location or explicit contractual consent. This concept aligns with models used in data protection laws such as the European Union's General Data Protection Regulation (GDPR), which applies extraterritorially based on the location of the data subject [77]. In addition, automated compliance checks and blockchain-based smart contracts can facilitate technical enforcement while enabling scalable and transparent regulation, contributing to broader systemic resilience and governance in metaverse ecosystems [78].

2) MISINFORMATION AND FAKE CONTENT

The ability of AIGC to produce realistic text, audio, images, and videos raises serious concerns around the spread of misinformation, fake news, deep fakes, and other manipulated content. In a highly immersive metaverse environments, malicious actors can exploit such content to deceive users, and manipulate public opinion. It can also be used to misrepresent virtual assets or investments, creating financial and reputational risks [79]. Furthermore, these risks are amplified by the emotional intensity and immediacy of metaverse interactions. Addressing this challenge requires both technical solutions, such as real-time content authentication and robust detection mechanisms, and regulatory measures to ensure accountability for malicious content creation and distribution.

3) PRIVACY AND DATA SECURITY

Metaverse applications powered by GenAI inherently rely on the continuous collection and processing of fine-grained personal data, including behavioral patterns, biometric signals, and interaction histories. This raises significant privacy and security concerns, particularly as GenAI models may inadvertently memorize or leak sensitive information [80]. Existing privacy regulations, including the GDPR [81] and the California Consumer Privacy Act (CCPA), provide useful legal guidance, however, applying these frameworks uniformly across decentralized and cross-border metaverse platforms remains a significant challenge [82]. To this end, technological solutions such as decentralized identity solutions and Privacy Enhancing Technologies (PET) such as differential privacy, can be leveraged to mitigate the privacy risks and preserve user autonomy.

4) TRAINING DATA BIAS AND DATA POISONING

AIGC systems trained on large-scale and real-world data are susceptible to perpetuating existing social biases and may generate outputs that are discriminatory or offensive. Furthermore, these systems are vulnerable to data poisoning attacks, where malicious actors inject misleading or harmful content into the training datasets, leading to systematically biased or harmful content creation. Addressing these challenges requires a combination of bias mitigation techniques (such as fairness-aware training and debiasing algorithms) and robust adversarial defenses to safeguard both the integrity of the models and the well-being of users [2].

5) INTELLECTUAL PROPERTY RIGHTS (IPRS)

The rise of AI-generated content challenges traditional notions of intellectual property, which largely presuppose human authorship. Current legal frameworks offer limited clarity on the ownership of works produced autonomously by AI, creating uncertainty for creators, platform operators, and users alike [83]. Potential solutions include the development of AI-assisted authorship frameworks, where ownership is determined based on the degree of human input, or the introduction of default public domain licenses for purely AI-generated works. Both approaches could help balance the interests of creators, AI developers, and the public while supporting innovation in the metaverse.

TABLE 3. Summary of Various AIGC Detection Techniques by Technique, Proposed Method, and Dataset. Methods are Categorized by Content Type: text, image, and video.

Work	Content Type	Detection Type	Datasets Used	Detection Method
GPTZero [89]	Text	Statistical-based		
GLTR [90]	Text	Statistical-based	BERT [91], GPT-2 [92], The bAbI task children book corpus (CBT)	Words Probability, Absolute Word Rank, Predicted Distribution Entropy
DetectGPT [93]	Text	Statistical-based	SQuAD WritingPrompts, XSum	Probability Curvature
OpenAI	Text	Classification-based		
Ma et al. [94]	Text	Classification-based	A collection of human-written abstracts and AI-generated abstracts	Features including writing style, coherence, consistency, argument logistics)
Guo et al. [95]	Text	Classification-based	Human ChatGPT Comparison Corpus (HC3)	A deep classifier based on a pre-trained LM
Kirchenbauer et al. [96]	Text	Classification-based	BOOKCORPUS, CC-NEWS, OPENWEBTEXT, STORIES	Pre-trained Transformer-based neural network fine-tuned for classification
RoBERTa [97]	Text	Watermarking-based	Colossal Clean Crawled Corpus (C4)	Exact Match (EM), F1 score (for watermark detection)
Minder et al. [98]	Text	Classification-based	A corpus of nearly 500 articles covering 10 topics (the Human-AI-Generated Text Corpus)	Classification using nine feature categories (perplexity, semantic, list lookup, document-level, error-based, readability, AI feedback, text vectors, etc.)
Yang et al. [99]	Image	Facial Feature-based	Faceforensics++	Image saliency detecting manipulated images
DA-FDFtNet [100]	Image	Facial Feature-based	Faceforensics++, PGGAN, StyleGAN, StyleGAN2	Pre-trained model, Fine-Tuned Transformer, MBblockV3, channel attention module
Faceforensics++ [101]	Image	Facial Feature-based	Faceforensics++	Face tracking by Thies et al. [102].
Suganthi et al. [103]	Image	Facial Feature-based	FFHQ, 100K-Faces, DFFD, CASIA-WebFace	DL of fisherface using Local Binary Pattern Histogram
Hsu et al. [104]	Image	Facial Feature-based	CelebA, DCGAN, WGAP, WGAN-GP, LSGAN, PGGAN	Contrastive loss to learn common discriminative features
Hsu et al. [105]	Image	Facial Feature-based	CelebA, DCGAN, WGAP, WGAN-GP, LSGAN, PGGAN	Contrastive loss and a two-stream network to encode pairwise information
Xi et al. [106]	Image	Generic Feature-based	Constructed Text-to-Image (T2I) datasets with DALL·E2 and DreamStudio	A residual stream and content stream to explore generation traces of T2I
Narvaez et al. [107]	Image	Generic Feature-based	ImageNet, ILSVRC-2012-CLS	CNN trained with an artistic styles dataset.
Guarnera et al. [108]	Image	Generic Feature-based	CelebA, RaFD	Expectation Maximization (EM) to detect forensic traces
Agarwal et al. [109]	Video	Visual Feature-based	VoxCeleb2	Users appearance and behavioral biometrics
In Ictu Oculi [110]	Video	Visual Feature-based	Constructed Eye-blinking Video Data	Eye-blinking detection
Deepfakestack [111]	Video	Visual Feature-based	FaceForensics++	Meta-learner trained on base-learners's outputs
Li & Lyu [112]	Video	Spatio-temporal Feature-based	UADFV, DeepfakeTIMIT	Detecting affine wrapping artifacts in DeepFake-synthesized face images
Lips don't lie [113]	Video	Spatio-temporal Feature-based	Lipreading in the Wild (LRW), FaceForensics++, Celeb-DF-v2	Learning internal representations of natural mouth motion.
[114]	Video	Spatio-temporal Feature-based	DFDC, FaceForensics++, Deeper Forensics	3D CNNs to learn deepfake-specific attributes.
Singh et al. [115]	Video	Spatio-temporal Feature-based	DFDC	Learns spatiotemporal features from frame sequences
Multimodaltrace [116]	Video	Audiovisual-based	FakeAVCeleb, World Leaders Dataset(WLD), Presidential Deepfake Dataset (PDD)	Fuses learned channels from audio/visual modalities via IntrAmodality Mixer Layer (IAML) and IntErModality Mixer Layer (IEML).

6) REGULATORY COMPLIANCE

Ensuring compliance with diverse regulatory guidelines is essential for building user trust in the metaverse. Legal frameworks such as the GDPR [84] and CCPA [85] emphasize transparency, user consent, and accountability, but applying these principles to immersive, fast-evolving environments poses significant challenges. Given the need to coordinate compliance across multiple stakeholders such as platform providers, content creators, AI service vendors, and end users, requires the development of interoperable compliance protocols and the embedding of accountability-by-design features into metaverse systems. This may involve dynamic consent management, audit trails for AI decisions, and real-time compliance monitoring.

IV. TOWARDS DETECTING AI-GENERATED CONTENT

In this section, we review existing methods for detecting AIGC and examine their robustness. We begin by outlining why AGIC detection is vital for the growing metaverse.

A. WHY DETECTING AIGC IS IMPORTANT IN METAVERSES?

The proliferation of AIGC has brought remarkable benefits across numerous domains—from entertainment to education—but it has also introduced substantial risks within increasingly immersive environments. Malicious actors can exploit AIGC to fabricate misleading or harmful content, jeopardizing the credibility and security of virtual spaces. Effective AIGC detection in the metaverse is essential for several reasons:

- *Ensuring credibility and Authenticity:* User trust depends on reliable, verifiable information. In metaverse settings such as social interactions, education, and commerce, the spread of deceptive AIGC undermines confidence and can lead to widespread misinformation.
- *Mitigating Amplified Risks:* Because the metaverse blurs the distinctions between real and artificial content, malicious creations can be more damaging and harder to recognize. Swift, scalable detection methods are crucial for identifying misleading AIGC in real time.
- *Keeping Pace with Evolving Threats:* As AIGC technologies advance, detection systems must evolve to address emerging manipulation tactics. Continuous research and innovation are critical for maintaining user safety and sustaining trust in metaverse platforms [86], [87], [88].

Given the capacity of AIGC to generate realistic text, images, and videos, there is a pressing need to prevent the dissemination of misinformation, fake news, and other deceptive material that can erode user trust. The virtual environment's low costs, ease of access, and rapid information sharing compound these challenges. In the following sections, we discuss a range of techniques for detecting diverse forms of AIGC. Table 3 summarizes notable AIGC detection methods by content type (text, image, video), detection approach, datasets used, and the underlying algorithms.

B. AI-GENERATED TEXTUAL CONTENT DETECTION

As AI-generated textual content grows in sophistication, substantial research has focused on verifying text authenticity.

Detection techniques generally fall into three main categories: statistical methods, classifier-based approaches, and watermarking strategies.

1) STATISTICAL DETECTION METHODS

These methods examine linguistic features such as entropy, perplexity, and n-gram frequency to uncover patterns typical of AI-generated text. For instance, Lavergne et al. [117] proposed an entropy-based detector that identifies semantic anomalies in AI-generated writing. Such approaches excel at spotting outliers that deviate from human linguistic norms.

2) CLASSIFIER-BASED DETECTION

Classifiers-based detection is common in tasks like fake news identification. Mindner et al. [98] utilized classifiers drawing on perplexity, semantic, readability, and vector-based features. Their method achieved F1 scores above 96% for purely AI-generated text and 78% for AI-rephrased text, outperforming tools like GPTZero [89] by 183.8% in F1 score. This underscores the importance of innovative features for robust detection systems.

3) WATERMARK-BASED DETECTION

Watermarking has gained traction as a covert labeling strategy for AI-generated text. Kirchenbauer et al. [97] introduced a watermarking approach that subtly marks text during sampling. Detection relies on a statistical test, requiring minimal tokens and no direct access to the underlying language model.

4) SPECIALIZED DETECTION TOOLS

Several specialized tools exist to detect AI-written text:

- *GPTZero:* Uses metrics like perplexity and burstiness to differentiate AI content from human-writing, under the assumption that human text typically exhibits greater variability.
- *DetectGPT:* Focuses on perturbations in model-generated text by comparing log probabilities of the original and modified versions [93].
- *HowkGPT:* Employs perplexity-based metrics to classify text originality [118].
- *OpenAI Classifier:* Fine-tuned models on paired datasets of human- and AI-generated text, although its 9% false positive rate underscores the need for further improvements [94].

5) EMERGING RESEARCH DIRECTIONS

Novel methods continue to surface for specific contexts. Guo et al. [95] employed RoBERTa-based classifiers tuned on the HC3 dataset for single-text and QA detection tasks. Authors also investigated syntactic and pragmatic cues to enhance the detection of AI-generated scientific writing [94].

C. AI-GENERATED VIDEO/IMAGE/AUDIO DEEPFAKE DETECTION

Deepfakes are synthetic or manipulated media (videos, images, audio) designed to portray individuals in fabricated scenarios. Powered by GANs, deepfake creation automates generative processes that produce strikingly realistic inauthentic content. To counter these threats, researchers have developed various ML methods to identify telltale inconsistencies in deepfake media [119], [120].

1) DEEPFAKE IMAGE DETECTION

As image manipulation tools proliferate, research has focused on detecting altered facial images. Approaches include:

a) Guided Filtering Yang et al. [99] harness guided filtering to amplify textural discrepancies in manipulated faces, enabling a more accurate differentiation of genuine vs. tampered images.

b) Transformer-Based Models: Bang et al. [100] introduced DA-FDFtNet, which integrates a fine-tuned transformer, MBblockV3, and a channel attention module for detecting manipulated faces.

c) CNN-Based Techniques: Rossler et al. [101] used CNNs trained on FaceForensics++ (covering both graphics- and learning-based manipulations). Suganthi et al. [103] leveraged Local Binary Pattern Histogram (LBPH) and deep belief networks for identifying fake face images.

d) Contrastive Learning: Hsu et al. [104], [105] proposed Deep Forgery Discriminator (DeepFD) method employing contrastive loss to detect subtle forged features in images, achieving up to 94.7% accuracy.

e) Forensic Signature Detection: Guarnera et al. [108] identified a "forensic signature" formed by the generative process, allowing detection of synthetic images via an Expectation Maximization (EM) algorithm.

2) DEEPFAKE VIDEO DETECTION

Videos present added complexity, requiring the analysis of both spatial and temporal cues.

a) Frame-Level Analysis: Guera et al. [121] combined CNN-based feature extraction on individual frames with RNNs to detect temporal inconsistencies from face swapping. Singh et al. [115] exploited spatiotemporal cues across consecutive frames.

b) Ensemble Methods: Rana et al. [111] developed DeepfakeStack, an ensemble meta-learner framework for video-level tampering detection.

c) Behavioral Biometrics: Agarwal et al. [109] applied CNNs with metric-learning to spot mismatched facial biometric in video segments, revealing inconsistencies in face-swapped clips.

d) Low-Level Anomalies: Li et al. [112] exposed artifacts from affine transformations during deepfake creation, while Haliassos et al. [113] emphasized mouth-movement anomalies for detection.

e) Spatio-Temporal Features: 3D CNNs have also been used to learn joint spatial-temporal representations, aiding generalization across diverse deepfake generation techniques [114].

f) Eye Blinking Cues: Li et al. [110] leveraged irregular eye-blinking patterns, training an LRCN model to detect blinking anomalies.

g) Multimodal Analysis: Raza et al. [116] fused audio and visual features via intrAmodality and interModality mixers to achieve up to 98% detection accuracy.

3) DETECTION OF OTHER AI-GENERATED MEDIA

Beyond human faces manipulation, researchers have addressed alternative AI-generated content types.

a) Text-to-Image (T2I) Detection: Xi et al. [106] proposed a dual-stream model with residual and content streams for differentiating T2I images from photographs.

b) AI-Generated Art: Narvaez et al. [107] trained deep models on stylistic and compositional features to separate human-made artwork from machine-generated art, exceeding 90% accuracy.

D. DETECTING AIGC IN THE METAVERSE

1) CAPABILITIES AND RISKS OF AIGC IN THE METAVERSE

GenAI models excel at producing highly realistic content in multiple modalities, raising concerns about misinformation, fake news, and manipulated media. Because the metaverse emphasizes realism and immersion, the authenticity of shared information can be severely compromised by synthetic content, undermining user trust and creating uncertainty in digital interactions.

2) LIMITATIONS OF EXISTING DETECTION METHODS

Most current detection systems rely on pre-trained models tailored to specific content types, such as fake news or manipulated images. These approaches often struggle with the metaverse's complexity, which features interactive 3D objects, real-time events, and AI-generated avatars. Consequently, existing methods must evolve to accommodate emerging AIGC paradigms.

3) EMERGING CHALLENGES IN AIGC DETECTION

The metaverse seamlessly merges real-time interactions, three-dimensional spaces, and multimodal content—making malicious content detection more complicated than in traditional online environments. New approaches must be developed to handle these unique content types and preserve user trust.

4) PROPOSED SOLUTIONS AND RESEARCH DIRECTIONS

Future research should investigate advanced techniques specifically crafted for metaverse settings. Lin et al. [122] proposed integrating blockchain and zero-knowledge proofs into semantic communication to differentiate adversarial from legitimate data. Such novel strategies can bolster the metaverse's defense against misleading AIGC. Beyond technical enhancements, blockchain can also contribute to the creation of new economic models and digital value systems in the metaverse, fostering decentralization, trust, and financial inclusion [123].

5) ANALYZING AIGC DETECTION TECHNIQUES IN METAVERSE SETTINGS

To address the unique challenges posed by the metaverse such as real-time content generation, immersive multi-modal environments, and resource constraints, it is essential to critically evaluate existing AIGC detection techniques beyond their general capabilities. It is worth noting that not all methods described above are equally suited for deployment in dynamic virtual environments where detection decisions may need to occur within milliseconds without disrupting user experience. Table 4 presents a qualitative comparison of prominent AIGC detection approaches, focusing on their strengths, limitations, and suitability for metaverse use cases. Statistical methods, though lightweight and interpretable, often fall short in accuracy against sophisticated AIGC and are susceptible to evasion attacks. On the other hand, classifier-based models such as ML/DL-based methods are capable of high accuracy, however, they typically require significant computation and fine-tuning that limits their use in real-time applications. Whereas, watermarking approaches can provide fast verification but depend on control over the content generation pipeline. Similarly, methods that focus on multimodal fusion can provide higher resilience, but they will also incur high complexity and latency. For better understanding, consider a metaverse educational platform where AI-driven non-player characters (NPCs) act as interactive tutors, delivering real-time instructional content, answering student questions, and facilitating group discussions. In this scenario, the authenticity of content is crucial to avoid misinformation, but detection systems must operate seamlessly to preserve engagement and avoid perceptible delays. The selection of detection methods would therefore depend on the specific nature of the interaction and the associated risk tolerance:

- *Statistical methods* can be deployed on-device to provide instantaneous and low-latency filtering of simple AI-generated responses.
- *Classifier-based approaches* such as DL-based techniques can be applied selectively for post-event analysis or asynchronous content moderation where response time is less critical but higher accuracy is required.
- *Watermark-based detection* can be integrated into the NPC content generation pipeline, enabling fast, lightweight authenticity checks for scripted or platform-generated AI responses without heavy computation.
- *Multimodal detection* (e.g., fusing audio tone, facial expressions, and text coherence) could be reserved for high-stakes educational simulations, such as exams, assessments, or sensitive topics, where the risk of

TABLE 4. Comparison of AIGC Detection Methods in the Context of Metaverse Applications

Method	Strengths	Limitations	Metaverse Applicability
Statistical Methods	Lightweight, interpretable, computationally efficient	Limited accuracy against advanced AIGC and AI models; vulnerable to paraphrasing	Suitable for low-latency, low-risk applications such as basic content filtering or simple NPC interactions
Classifier-Based Detection	Can achieve high detection accuracy in narrow or well-defined contexts	Requires fine-tuning; poor generalization to unseen models or content; high computational costs	Often impractical for real-time detection; better suited for offline moderation
Watermark-Based Detection	Efficient runtime detection; resistant to common manipulations; fast verification	Requires control over generation pipeline; ineffective for third-party or legacy content; raises ethical concerns	Well-suited for controlled generation settings such as scripted NPCs or platform-owned AI
Multimodal Fusion Detection	High robustness to manipulation; leverages cross-modal cues for enhanced accuracy	Computationally intensive; complex system integration; potential latency	Applicable in high-stakes or security-sensitive metaverse scenarios requiring stronger assurance (e.g., finance, governance)

deception or manipulation requires additional scrutiny despite increased resource costs.

This case study highlights that no single detection technique is universally optimal for the metaverse applications that are highly immersive and require low latency. Instead, adaptive, context-aware selection of detection strategies can be employed, thereby balancing latency, accuracy, and resource consumption, which is essential to maintain both security and seamless user experience.

E. ROBUSTNESS OF AIGC DETECTORS

Although numerous AIGC detectors have been developed for text, images, and video, it remains critical to evaluate their resilience against adversarial manipulations. Common performance metrics—accuracy, precision, recall, and F1 score—are used to measure a detector's ability to consistently identify or reject adversarial examples.

1) EVADING TEXT DETECTION

The task of detecting AI-generated text is especially challenging given how easily adversaries can manipulate semantics, style, and other subtleties. Cai and Cui [124] highlighted how small formatting tweaks (e.g., extra spaces) can considerably degrade detector performance. They introduced SpaceInfi, an approach that appends spaces before commas, effectively bypassing many state-of-the-art detectors. Zou et al. [125] described a lightweight strategy that appends specific suffixes to prompts, prompting LLMs to generate content that evades detection. Their method is especially potent against GPT-based models. Krishna et al. [126] used DIPPER, an 11-billion-parameter paraphraser, to reword AI-generated text, reducing detection accuracy for various tools (GPTZero, DetectGPT, OpenAI's classifier) from as high as 70% to under 5%. Liang et al. [127] revealed how most detectors disproportionately misclassify non-English text as AI-generated while correctly identifying native English texts. Prompting strategies that mimic English writing patterns effectively bypass these detectors. Wang et al. [87] highlighted the difficulty of identifying AI-generated code, calling for more specialized methods. Flores et al. [128] introduced a benchmark focusing on compositional semantics, lexical relationships, and sensitivity to modifiers. Their findings underscore the struggle of current detectors with nuanced textual manipulations.

2) EVADING DEEPFAKE (IMAGE/VIDEO/AUDIO) DETECTION

DeepFake detection relies primarily on classifying manipulated images or video frames. However, adversarial examples

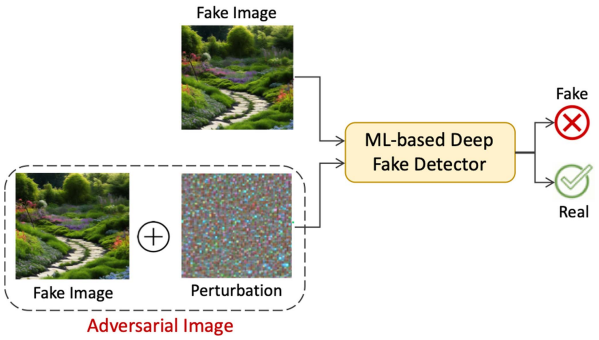

FIGURE 5. Illustration of how adversarial perturbations (often imperceptible) can defeat ML-based DeepFake detectors.

(see Fig. 5) can degrade the performance of many ML–based detectors with minimal impact on perceived content quality. In the literature, various studies have focused on developing attacks to evade DeepFake detectors. For instance, Liao et al. [129] selectively perturbed crucial pixels, achieving state-of-the-art results with minimal L_0 and L_2 distortions compared to standard attacks. Hussain et al. [130] showed that perturbing individual frames in fake videos can mislead detectors into labeling them as real, revealing a major vulnerability in most detection systems. Similarly, Gandhi and Jain [131] exploited techniques such as Fast Gradient Sign Method (FGSM) to reduce detection accuracy from over 95% to below 27% in certain scenarios. The authors in [132], studied the robustness of sequence-based detectors, demonstrating successful attack rates exceeding 99% under white-box conditions. A few studies have also investigated the generalizability and transferability of adversarial perturbations. For example, Neekhara et al. [133] introduced universal perturbations that generalize across multiple fake video frames and different detection architectures, where as Vo et al. [134] highlighted the ease of transferring white-box attacks to black-box models, underscoring the need for integrated defense strategies. Fig. 6 presents a taxonomy of AIGC detection techniques and evasion strategies, illustrating current approaches and gaps for further research.

3) BALANCING FIDELITY AND SECURITY IN AIGC

While enhancing the fidelity and personalization of AIGC is essential to create immersive metaverse experiences, it also raises important security challenges. In particular, increasing the realism of synthetic content such as hyper-realistic avatars, personalized voice synthesis, or lifelike gestures

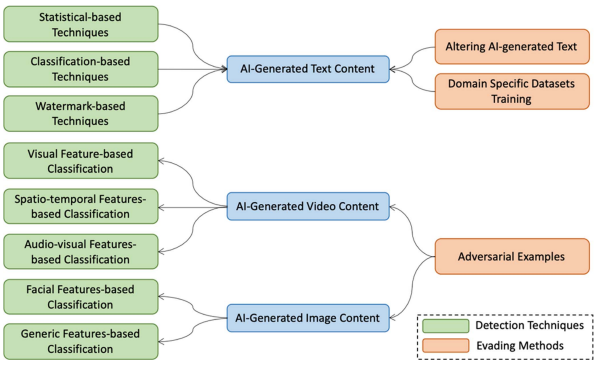

FIGURE 6. Taxonomy of AIGC detection techniques and evasion methods, categorized by content type.

may inadvertently undermine the effectiveness of existing detection systems. Specifically, methods that rely on low-level artifacts or statistical features may struggle to differentiate between authentic and synthetic content, particularly as generative models improve in sophistication [135]. This amplifies the challenges related to misinformation, deepfake detection, and content authentication. To address this issue, we advocate for adopting a security-by-design approach in AIGC development. Such an approach involves embedding detection compatibility and provenance mechanisms directly into content generation workflows from the outset, rather than relying solely on post hoc defensive measures. This principle is consistent with well-established practices in secure software engineering, which emphasize proactive, systemic risk mitigation over reactive patching [136].

F. EVALUATION METRICS

Various metrics are employed to gauge the performance of AIGC detection methods, including Accuracy, Precision, Recall, and F1 Score.

- *Accuracy:* Proportion of correctly classified cases among all instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

  where TP, TN, FP, and FN represent True Positives, True Negatives, False Positives, and False Negatives, respectively.
- *Precision:* Fraction of true positives among predicted positives (crucial for avoiding false alarms).

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

- *Recall (Sensitivity):* Fraction of true positives detected among all actual positives (important to minimize missed detections).

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

- *F1 Score:* Harmonic mean of Precision and Recall, balancing their trade-offs.

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

When detecting AIGC, achieving high Precision helps avoid mistakenly flagging legitimate content as AI-generated, while high Recall ensures that genuinely AI-generated content does not slip through. The F1 Score offers a balanced gauge of overall detection capability—especially relevant for metaverse scenarios where AIGC threats evolve rapidly and exhibit diverse forms.

G. HUMAN-CENTERED PERSPECTIVES ON AIGC IN THE METAVERSE

End-user perceptions of AIGC, particularly regarding its authenticity, trustworthiness, and ethical implications, are crucial for ensuring positive engagement and psychological comfort within immersive metaverse environments. The relevant literature highlights that user trust in AI plays a decisive role in shaping human–AI interactions [137]. Similarly, Shin et al. [138] demonstrated that users' ability to understand and rationalize AI behavior is critical for promoting perceptions of fairness and ethical alignment, which subsequently strengthens trust and acceptance. Moreover, Choung et al. [139] identified two distinct dimensions of user trust in AI technologies, i.e., human-like trust and functionality trust. Their findings indicate that while both trust dimensions influence users' attitudes and intention to adopt AI, functionality trust exerts a stronger overall impact on behavioral intention. In [140], the authors empirically demonstrated that both cognitive trust, influenced by users' perceptions of AI's intelligence, transparency, and accuracy, and affective trust, shaped by perceptions of AI empathy, are essential for building overall trust in AIGC. Their findings revealed that trust in AIGC significantly predicts users' adoption intentions. Moreover, they found that algorithmic bias weakens the relationship between cognitive trust and trust in AIGC, underscoring the need for improved fairness and transparency in AI content delivery to sustain user trust.

V. OPEN RESEARCH ISSUES

The adoption of GenAI in the metaverse raises diverse technical, ethical, governance, and societal challenges. This section presents the key open research issues into three thematic categories based on their relative urgency: (1) advancing security and robustness to counter threats such as misinformation and deepfakes, (2) ensuring ethical, inclusive, and responsible AI development for social acceptance and trust, and (3) advancing future AIGC capabilities through immersive AI innovation and systematic validation.

A. ADVANCING SECURITY AND ROBUSTNESS OF AIGC IN METAVERSE

1) ADVANCING ROBUST AND ADAPTIVE AIGC DETECTION TECHNIQUES

The metaverse's varied and ever-evolving AIGC landscape demands adaptable, scalable, and accurate detection frameworks:

- *Generalizable Detection Frameworks:* Multi-modal detection approaches that simultaneously analyze text, images, videos, and 3D content are essential for handling complex metaverse environments and emerging content types.
- *Enhancing DeepFake Detection:* Strengthening DeepFake detection systems calls for standardized data preprocessing, improved resilience against adversarial attacks, and adaptable models that can handle multiple DeepFake formats (e.g., face swapping, motion manipulation).
- *Resilience Against Detection Evasion:* Adversarial scenarios unique to the metaverse—such as texture perturbations in 3D worlds or paraphrasing in textual content—require sophisticated, context-aware defenses. Collaborative research efforts can lead to standardized, robust solutions.

By addressing these interconnected challenges through robust frameworks, adaptive systems, and targeted detection enhancements, AIGC detection in the metaverse can achieve the scalability, reliability, and security needed for its dynamic environments.

2) ADDRESSING COMPUTATIONAL BOTTLENECKS AND HARDWARE CHALLENGES

Real-time AIGC generation and detection in resource-intensive metaverse environments present significant computational challenges, particularly when considering the stringent latency, scalability, and hardware constraints. To meet these demands while maintaining an immersive user experience, following complementary strategies can be considered:

- *Hybrid Edge-Cloud Architectures:* Splitting computational tasks between edge devices and the cloud can optimize resource utilization and reduce latency. In this approach, latency-sensitive operations such as initial AIGC moderation for NPC dialogue or real-time virtual interactions can be performed on edge devices, while computationally intensive tasks such as large models-based inferences or anomaly detection can be offloaded to cloud servers. For this purpose, techniques such as split inference allow portions of large models to run on-device with minimal delay, while heavier computations complete asynchronously in the cloud [141].
- *Optimizing Hardware Performance:* Enhancing hardware efficiency through the use of specialized processors (e.g., GPUs, TPUs, etc.) combined with model optimization techniques such as pruning [142], quantization [143], and knowledge distillation [144], can significantly reduce computational and energy overhead without sacrificing detection accuracy. Such lightweight models are particularly suited for XR headsets and handheld devices where power and thermal limitations are critical.
- *Scalable and Adaptive Architectures:* Metaverse environments are characterized by highly dynamic user populations. To ensure scalability, AIGC detection and generation systems should be designed using microservice-based architectures [145], which support autoscaling and dynamic load balancing. In addition, adaptive fidelity control, similar to adaptive bitrate streaming can be employed to modulate the complexity or richness of AI-generated content based on available system resources and network conditions [146].

3) ENHANCING ROBUSTNESS AND VALIDATING DETECTION TECHNIQUES IN THE METAVERSE

Ensuring that AIGC detection methods operate reliably under the metaverse's high-frequency, interactive conditions is paramount:

- *Multi-Modal and Context-Aware Detection:* Detection systems must seamlessly handle text, audio, images, video, and 3D content. Adversarial threats (e.g., 3D texture manipulations or temporal distortions in video conferencing) demand context-sensitive adaptive defenses.
- *Integration with User-Generated Content:* AI-generated assets must harmonize with user-generated materials to maintain immersive experiences. Context-aware systems capable of on-the-fly adaptation can better align with users' real-time inputs and environmental cues.
- *Real-World Validation:* Benchmarks mirroring authentic metaverse scenarios—gaming, remote training, collaborative design—are necessary to examine scalability, reliability, and usability. Standardized metrics and controlled experiments will further sharpen detection research.

B. ENABLING ETHICAL, INCLUSIVE, AND RESPONSIBLE AIGC DEPLOYMENT

1) ETHICAL DESIGN FRAMEWORKS

Embedding ethical principles into the metaverse development lifecycle is crucial for fostering safe, responsible user experiences:

- textitFairness, Accountability, Transparency: Existing guidelines (e.g., IEEE's Ethically Aligned Design [147], EU Commission's trustworthy AI principles [148]) provide foundational concepts, but new frameworks are needed to address metaverse-specific challenges.
- *Bias Mitigation:* Careful oversight is required to identify and reduce biases in AI-driven processes, preventing the reinforcement of harmful stereotypes or discriminatory outcomes.

- *Adaptive Governance:* Ethical guidelines should evolve with advances in metaverse technologies to remain relevant and effective.

2) ENHANCING USER PRIVACY AND DATA PROTECTION

The immersive character of the metaverse necessitates collecting and processing detailed user data, raising serious privacy issues:

- *User Control:* Users must be well-informed about how their data is gathered, stored, and utilized, with clear consent and opt-out options.
- *Privacy-Preserving Techniques:* Methods like encryption, differential privacy, and anonymization can safeguard sensitive data [149].
- *Transparent Data Practices:* Clear, accessible disclosures about data flows and regulatory compliance (e.g., GDPR) help establish trust among users.

3) DECENTRALIZED GOVERNANCE FOR THE METAVERSE

Decentralization offers potential for inclusive, community-driven decision-making in the metaverse:

- *Decentralized Autonomous Organizations (DAOs):* DAOs allow shared governance of virtual spaces, distributing power among stakeholders, and can be instrumental in community-driven policy enforcement [150].
- *Scalability and Fair Representation:* Ongoing research must address how to scale governance structures while maintaining equitable representation and preventing power concentration.

4) MULTILINGUAL AND EQUITABLE LLMS FOR THE METAVERSE

Current LLMs predominantly rely on English-language training data, limiting their effectiveness for non-English speakers and raising concerns about inclusivity in a globally oriented metaverse. While improving multilingual capabilities is essential, equitable participation in the metaverse also depends on addressing broader systemic inequalities beyond language alone. Therefore, research efforts should not only be devoted to enhancing language coverage but also to accounting for cultural, economic, infrastructural, and epistemological disparities that may shape users' ability to access and benefit from metaverse technologies. For instance, even with multilingual support, users from low-resource regions may face barriers due to limited connectivity, lack of affordable devices, or cultural misalignment with AI-generated content. Beyond these barriers, there is also a risk of cultural homogenization where dominant language groups, Western values, and mainstream digital norms marginalize indigenous knowledge systems, non-Western worldviews, and locally meaningful cultural expressions. For instance, AI models trained without local input risk not only linguistic exclusion but also the erasure or distortion of cultural narratives, as demonstrated in African contexts where LLMs fabricated culturally inaccurate information [151]. Moreover, the recent literature suggests that data-driven systems, including language models, risk perpetuating colonial dynamics by privileging certain ways of knowing while erasing others [152], [153]. To support equitable metaverse access, future research and development should consider the following:

- *Language Coverage:* Adapting LLMs or developing specialized language models to serve non-English speakers and reflect the full diversity of linguistic expressions, including low-resource and indigenous languages.
- *Cultural Sensitivity:* Incorporating culturally aware training data and ethical guidelines to ensure that AI-generated content respects diverse norms, values, and identities.
- *Addressing Biases:* Mitigating linguistic and cultural biases to ensure that multilingual LLMs foster inclusive metaverse interactions and represent diverse user backgrounds fairly.
- *Accessibility and Affordability:* Designing lightweight models and decentralized architectures to make GenAI tools accessible in low-resource settings, addressing economic and infrastructural inequalities.
- *Decolonial Pluralism:* Embedding epistemic plurality by including Indigenous knowledge systems, non-Western cultural narratives, and locally relevant value frameworks into LLM training and content generation. This involves not only multilingualism but also respect for diverse worldviews and modes of expression to ensure the metaverse reflects the full spectrum of human cultures and not just dominant narratives.
- *Benchmarking Inclusivity:* Establishing standardized multilingual and multicultural benchmarks to assess not only language performance but also fairness, representation, and cultural relevance.

C. ADVANCING AIGC AND DETECTION CAPABILITIES IN THE METAVERSE

1) FUTURE OF GENAI IN THE METAVERSE

Progress in GenAI for the metaverse revolves around three main areas:

- *Customization and Quality:* Continued research is important for refining models, algorithms, and architectures that produce high-fidelity, personalized content, making AI creations indistinguishable from human-generated materials.
- *Applications in Metaverse Content Creation:* GenAI can substantially streamline content generation for applications like education, design, and commerce, including the automated creation of virtual worlds, objects, characters, and user-driven experiences.
- *Integration with Emerging Technologies:* Combining GenAI with AR/VR, blockchain, and Internet of Things (IoT) opens new avenues. Blockchain, for instance, can enhance content ownership and facilitate secure AIGC communication in the metaverse [122], [154].

2) TOWARDS ROBUST AND SCALABLE AIGC DETECTION

While the conventional evaluation metrics provide valuable insights into detection effectiveness, but do not fully capture the practical complexities of real-time metaverse applications where speed, scalability, and responsiveness are equally important. Unlike current social networking platforms, the metaverse presents unique challenges for AIGC detection due to its inherently dynamic, low-latency, and multi-modal nature. For instance, the metaverse blends text, audio, images, video, and 3D interactions simultaneously, all of which must be processed and analyzed in near real-time to maintain immersion and usability. In such highly dynamic and multi-modal settings, evasion strategies such as paraphrasing in text, pixel-level perturbations in images, or in videos become harder to detect at scale. Even subtle detection delays can undermine the security and safety of users. Therefore, deploying robust AIGC detectors in the metaverse requires balancing detection accuracy with computational efficiency. Models that are highly robust to adversarial attacks often incur significant computational overhead, which may not be feasible for devices operating under strict latency and resource constraints. On the other hand, lightweight detection models may offer limited resilience to sophisticated evasion strategies. Addressing these challenges will require advances in multi-modal detection algorithms that are both computationally efficient and resilient to adversarial manipulation. Achieving robust AIGC detection in the metaverse involves careful system design choices that align detection fidelity with latency requirements and resource availability. This remains an open research challenge requiring cross-disciplinary innovation at the intersection of AI security, systems design, and real-time computing. In large-scale deployments involving millions of concurrent users and diverse modalities, even minor adversarial vulnerabilities can propagate rapidly across networked interactions, thus amplifying risks and complicating containment.

3) ROADMAP FOR FUTURE VALIDATION OF AIGC DETECTION TECHNIQUES

Bridging the gap between theoretical advances and real-world applicability requires systematic validation:

- *Developing Benchmark Datasets:* Comprehensive, open-source datasets that reflect the metaverse's multimodal, interactive nature will enable consistent evaluations of detection methods.
- *Evaluating Robustness Under Adversarial Conditions:* Rigorous testing against adversarial manipulations—ranging from simple input perturbations to advanced deepfake techniques—will help refine and strengthen detectors.
- *Controlled Experiments in Virtual Environments:* Testing in simulated metaverse scenarios (e.g., virtual meetings, gaming, educational applications) can yield insights into scalability, user experience, and system performance.
- *Standardizing Evaluation Metrics:* Consistent metrics facilitate fair comparisons and accelerate progress in developing effective detection solutions.
- *Collaborative Validation Efforts:* Joint initiatives among academics, industry, and policy entities will ensure alignment between emerging detection frameworks and real-world metaverse demands.

VI. CONCLUSION

Generative AI technologies have opened up vast opportunities for content creation and user engagement across diverse domains, with the metaverse serving as a compelling next frontier for AIGC. However, concerns related to security, privacy, authenticity, and responsible use call for immediate scholarly and practical attention. In this paper, we provide a comprehensive exploration of AIGC fundamentals, the range of enabling GenAI technologies, and their potential in shaping metaverse environments. We also categorize existing detection approaches aimed at curbing malicious or misleading content, assessing their robustness against adversarial manipulations. By identifying critical gaps and presenting open research questions, this work offers a roadmap for advancing secure, resilient, and ethically grounded AIGC frameworks in metaverse platforms. Through continued innovation and collaboration, stakeholders can harness the transformative power of GenAI while safeguarding users' trust and well-being in these evolving virtual worlds.

REFERENCES

[1] J. Sun, W. Gan, H.-C. Chao, and P. S. Yu, "Metaverse: Survey, applications, security, and opportunities," 2022, *arXiv:2210.07990*.

[2] Y. Wang et al., "A survey on metaverse: Fundamentals, security, and privacy," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 319–352, Firstquarter 2023.

[3] N. Anantrasirichai and D. Bull, "Artificial intelligence in the creative industries: A review," *Artif. Intell. Rev.*, vol. 55, pp. 589–656, 2022.

[4] H. Ning et al., "A survey on the metaverse: The state-of-the-art, technologies, applications, and challenges," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14671–14688, Aug. 2023.

[5] C. Zhang et al., "One small step for generative AI, one giant leap for AGI: A complete survey on ChatGPT in AIGC era," 2023, *arXiv:2304.06488*.

[6] L.-H. Lee, P. Zhou, C. Zhang, and S. Hosio, "What if we have MetaGPT? From content singularity to human-metaverse interaction in AIGC era," 2023, *arXiv:2304.07521*.

[7] H. Wu, Y. Liao, R. Hadi Mogavi, P. Hui, and P. Y. Zhou, "Deepfake in the metaverse: An outlook survey," in *Proc. Int. Conf. Hum.-Comput. Interac.*, Cham: Springer, May 2025, pp. 253–267.

[8] S. Tariq, A. Abuadbba, and K. Moore, "Deepfake in the metaverse: Security implications for virtual gaming, meetings, and offices," in *Proc. 2nd Workshop Secur. Implic. Deepfakes Cheapfakes*, Jul. 2023, pp. 16–19.

[9] J. Wu, W. Gan, Z. Chen, S. Wan, and H. Lin, "AI-generated content (AIGC): A survey," 2023, *arXiv:2304.06632*.

[10] L. Basyoni and J. Qadir, "AI generated content in the metaverse: Risks and mitigation strategies," in *Proc. 2023 Int. Symp. Netw., Comput. Commun.*, 2023, pp. 1–4.

[11] X. Wang, Y. Hong, and X. He, "Exploring artificial intelligence generated content (AIGC) applications in the metaverse: Challenges, solutions, and future directions," *IET Blockchain*, vol. 4, pp. 365–378, 2024.

[12] M. Xu et al., "Sparks of generative pretrained transformers in edge intelligence for the metaverse: Caching and inference for mobile artificial intelligence-generated content services," *IEEE Veh. Technol. Mag.*, vol. 18, no. 4, pp. 35–44, Dec. 2023.

[13] X. Chen, L. Luo, F. Tang, M. Zhao, and N. Kato, "AIGC-based evolvable digital twin networks: A road to the intelligent metaverse," *IEEE Netw.*, vol. 38, no. 6, pp. 370–379, Nov. 2024.

[14] A. Qayyum et al., "Secure and trustworthy artificial intelligence-extended reality (AI-XR) for metaverses," *ACM Comput. Surv.*, vol. 56, pp. 1–38, 2023.

[15] A. W. Graylin and L. Rosenberg, *Our Next Reality: How the AI-Powered Metaverse Will Reshape the World*. Paris, France: Hachette, 2024.

[16] B. Ortutay, "Facebook Inc. rebrands as meta to stress 'metaverse' plan," Oct. 2021. [Online]. Available: https://www.denverpost.com/2021/10/28/facebook-meta-rebrand/#:~text=PUBLISHED%3A%20October%2028%2C%202021%20at,Zuckerberg%20calls%20the%20' 'metaverse

[17] M. Damar, "Metaverse shape of your life for future: A bibliometric snapshot," *J. Metaverse*, vol. 1, no. 1, pp. 1–8, 2021.

[18] F. Abdullakutty, A. Qayyum, and J. Qadir, "Trustworthy AI for educational metaverses," *Authorea Preprints*, 2024.

[19] A. Qayyum et al., "Can we revitalize interventional healthcare with AI-XR surgical metaverses?," in *Proc. 2023 IEEE Int. Conf. Metaverse Comput., Netw. Appl.*, 2023, pp. 496–503.

[20] A. Kumar, A. Shankar, A. Behl, B. B. Gupta, and S. Mavuri, "Lights, camera, metaverse!: Eliciting intention to use industrial metaverse, organizational agility, and firm performance," *J. Glob. Inf. Manage.*, vol. 31, no. 8, pp. 1–20, 2023.

[21] M. Deveci, D. Pamucar, I. Gokasar, M. Köppen, and B. B. Gupta, "Personal mobility in metaverse with autonomous vehicles using Q-rung orthopair fuzzy sets based OPA-RAFSI model," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 15642–15651, Dec. 2023.

[22] I. Gokasar, D. Pamucar, M. Deveci, B. B. Gupta, L. Martinez, and O. Castillo, "Metaverse integration alternatives of connected autonomous vehicles with self-powered sensors using fuzzy decision making model," *Inf. Sci.*, vol. 642, 2023, Art. no. 119192.

[23] M. Deveci, D. Pamucar, I. Gokasar, M. Köppen, B. B. Gupta, and T. Daim, "Evaluation of metaverse traffic safety implementations using fuzzy einstein based logarithmic methodology of additive weights and topsis method," *Technological Forecasting Social Change*, vol. 194, 2023, Art. no. 122681.

[24] W. Wider, L. Jiang, J. Lin, M. A. Fauzi, J. Li, and C. K. Chan, "Metaverse chronicles: A bibliometric analysis of its evolving landscape," *Int. J. Hum.–Comput. Interaction*, vol. 40, pp. 4873–4886, 2023.

[25] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.

[26] A. Ramesh et al., "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.

[27] E. Horvitz and T. M. Mitchell, "Scientific progress in artificial intelligence: History, status, and futures," in *Proc. Realizing Promise Minimizing Perils Artif. Intell. Sci. Community*, K. H. Jamieson, Ed., Philadelphia, PA, USA: University of Pennsylvania Press, 2024, p. 2024.

[28] M. Kang et al., "Scaling up GANs for text-to-image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 10124–10134.

[29] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 6840–6851.

[30] Z. Li, Z. Yang, and M. Wang, "Reinforcement learning with human feedback: Learning dynamic choices via pessimism," 2023, *arXiv:2305.18438*.

[31] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 4302–4310.

[32] A. Radford et al., "Improving language understanding by generative pre-training," 2018.

[33] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 5485–5551, 2020.

[34] 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," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[35] Z. Peng, Y. Dai, Q. Tang, X. Cui, and S. Guo, "Show and tell: A neural image caption generator," in *Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.

[36] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," 2019, *arXiv:1908.03557*.

[37] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole, "Zero-shot text-guided object generation with dream fields," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 857–866.

[38] N. Mohammad Khalid, T. Xie, E. Belilovsky, and T. Popa, "CLIP-Mesh: Generating textured meshes from text using pretrained image-text models," in *Proc. SIGGRAPH Asia 2022 Conf. Papers*, 2022, pp. 1–8.

[39] C. Zhang et al., "A complete survey on generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 all you need?," 2023, *arXiv:2303.11717*.

[40] D. Jurafsky and J. H. Martin, "Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition,".

[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[42] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights Imag.*, vol. 9, pp. 611–629, 2018.

[43] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, vol. 25. pp. 84–90.

[44] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. 13th Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 2014, pp. 818–833.

[45] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[46] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, vol. 27, pp. 2672–2680.

[47] C. Zhou et al., "A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT," in *Proc. Int. J. Mach. Learn. Cybern.*, 2023, pp. 1–65.

[48] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen, "Point-E: A system for generating 3D point clouds from complex prompts," 2022, *arXiv:2212.08751*.

[49] N. Corporation, "Nvidia avatar cloud engine (ACE) brings AI-powered avatars to games and virtual worlds," 2023. Accessed: Jul. 5, 2025. [Online]. Available: https://blogs.nvidia.com/blog/omniverse-ace-early-access/

[50] M. Zawish et al., "AI and 6G into the metaverse: Fundamentals, challenges and future research trends," *IEEE Open J. Commun. Soc.*, vol. 5, pp. 730–778, 2024.

[51] H. X. Qin and P. Hui, "Empowering the metaverse with generative AI: Survey and future directions," in *Proc. IEEE 43rd Int. Conf. Distrib. Comput. Syst. Workshops*, 2023, pp. 85–90.

[52] J. Ratican, J. Hutson, and A. Wright, "A proposed meta-reality immersive development pipeline: Generative AI models and extended reality (XR) content for the metaverse," *J. Intell. Learn. Syst. Appl.*, vol. 15, pp. 24–35, 2023.

[53] Y. Lin et al., "A unified framework for integrating semantic communication and AI-generated content in metaverse," *IEEE Netw.*, vol. 38, no. 4, pp. 174–181, 2023.

[54] J. Roberts, A. Banburski-Fahey, and J. Lanier, "Steps towards prompt-based creation of virtual worlds," 2022, *arXiv:2211.05875*.

[55] Z. Chen, G. Wang, and Z. Liu, "Text2Light: Zero-shot text-driven HDR panorama generation," *ACM Trans. Graph.*, vol. 41, no. 6, pp. 1–16, 2022.

[56] H. Huang, M. Solah, D. Li, and L.-F. Yu, "Audible panorama: Automatic spatial audio generation for panorama imagery," in *Proc. 2019 CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–11.

[57] M. Prpa, K. Tatar, J. Françoise, B. Riecke, T. Schiphorst, and P. Pasquier, "Attending to breath: Exploring how the cues in a virtual environment guide the attention to breath and shape the quality of experience to support mindfulness," in *Proc. 2018 Designing Interactive Syst. Conf.*, 2018, pp. 71–84.

[58] V. Chamola et al., "Beyond reality: The pivotal role of generative AI in the metaverse," 2023, *arXiv:2308.06272*.

[59] K. Nagano et al., "paGAN: Real-time avatars using dynamic textures," *ACM Trans. Graph.*, vol. 37, no. 6, 2018, Art. no. 258.

[60] D. Wang, P. Chandran, G. Zoss, D. Bradley, and P. Gotardo, "MoRF: Morphable radiance fields for multiview neural head modeling," in *Proc. ACM SIGGRAPH 2022 Conf.*, 2022, pp. 1–9.

[61] J. Yu, C. W. Chen, and Z. Wang, "3D singing head for music VR: Learning external and internal articulatory synchronicity from lyric, audio and notes," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 945–952.

[62] Z. Qiu et al., "SCULPTOR: Skeleton-consistent face creation using a learned parametric generator," *ACM Trans. Graph.*, vol. 41, no. 6, pp. 1–17, 2022.

[63] T. Randhavane, A. Bera, K. Kapsaskis, R. Sheth, K. Gray, and D. Manocha, "EVA: Generating emotional behavior of virtual agents using expressive features of gait and gaze," in *Proc. ACM Symp. Appl. Percep.*, 2019, pp. 1–10.

[64] G. C. Dobre, M. Gillies, and X. Pan, "Immersive machine learning for social attitude detection in virtual reality narrative games," *Virtual Reality*, vol. 26, no. 4, pp. 1519–1538, 2022.

[65] S. Conway and M. Ouellette, "Playing it cool: Considering Mcluhan's hot and cool taxonomy for game studies," *Convergence*, vol. 26, no. 5-6, pp. 1211–1225, 2020.

[66] G. Trichopoulos, G. Alexandridis, and G. Caridakis, "A survey on computational and emergent digital storytelling," *Heritage*, vol. 6, no. 2, pp. 1227–1263, 2023.

[67] G. Sivasankar, "Study of blockchain technology, AI and digital networking in metaverse," *IRE J.*, vol. 5, no. 8, pp. 110–115, 2022.

[68] J. H.-J. Choi and T. Chiang, "Living with soft dragons: Between science fiction and human-computer interaction," *Interactions*, vol. 29, no. 6, pp. 18–20, 2022.

[69] Z. Lyu, "State-of-the-art human-computer-interaction in metaverse," *Int. J. Hum.–Comput. Interaction*, vol. 40, pp. 6690–6708, 2023.

[70] Z. Lv, "Generative artificial intelligence in the metaverse era," *Cogn. Robot.*, vol. 3, pp. 208–217, 2023.

[71] R. M. Gil, D. Gutiérrez-Ujaque, and M. Teixidó, "Analyzing the metaverse: Computer games, blockchain, and 21st-century challenge," *Int. J. Hum.–Comput. Interaction*, vol. 40, pp. 6758–6775, 2023.

[72] Z. Cheng et al., "Cross-modal 3D shape generation and manipulation," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 303–321.

[73] K. Brodt and M. Bessmeltsev, "Sketch2pose: Estimating a 3D character pose from a bitmap sketch," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 1–15, 2022.

[74] Y. Sun, Y. Xu, C. Cheng, Y. Li, C. H. Lee, and A. Asadipour, "Travel with wander in the metaverse: An AI chatbot to visit the future earth," in *Proc. IEEE 24th Int. Workshop Multimedia Signal Process.*, 2022, pp. 1–6.

[75] A. J. Colangelo, "The legal limits of universal jurisdiction," *Virginia J. Int. Law*, vol. 47, 2006, Art. no. 149.

[76] H. X. Qin, Y. Wang, and P. Hui, "Identity, crimes, and law enforcement in the metaverse," *Humanities Social Sci. Commun.*, vol. 12, no. 1, pp. 1–15, 2025.

[77] C. Kuner, L. Bygrave, C. Docksey, and L. Drechsler, "The EU general data protection regulation: A commentary," *Update Sel. Articles*, 2020.

[78] P. Radanliev, "Review and comparison of USA, EU, and UK regulations on cyber risk/security of the current blockchain technologies: Viewpoint from 2023," *Rev. Socionetwork Strategies*, vol. 17, no. 2, pp. 105–129, 2023.

[79] K. Farish, "Do deepfakes pose a golden opportunity? Considering whether english law should adopt California's publicity right in the age of the deepfake," *J. Intellectual Property Law Pract.*, vol. 15, no. 1, pp. 40–48, 2020.

[80] C. Chen, Z. Wu, Y. Lai, W. Ou, T. Liao, and Z. Zheng, "Challenges and remedies to privacy and security in AIGC: Exploring the potential of privacy computing, blockchain, and beyond," 2023, *arXiv:2306.00419*.

[81] C. Kuner, L. A. Bygrave, C. Docksey, L. Drechsler, and L. Tosoni, "The EU general data protection regulation: A commentary/update of selected articles," *Update Sel. Articles*, 2021.

[82] Y. Wang, Y. Pan, M. Yan, Z. Su, and T. H. Luan, "A survey on ChatGPT: AI-generated contents, challenges, and solutions," *IEEE Open J. Comput. Soc.*, vol. 4, pp. 280–302, 2023.

[83] A. Ramalho, *Intellectual Property Protection for AI-Generated Creations: Europe, United States, Australia and Japan*. Evanston, IL, USA: Routledge, 2021.

[84] General data protection regulation, European Commission, 2024. Accessed: Jun. 10, 2024. [Online]. Available: https://gdpr-info.eu/

[85] California consumer privacy act (CCPA), California Attorney General's Office, 2024. Accessed: Jun. 10, 2024. [Online]. Available: https://oag.ca.gov/privacy/ccpa

[86] J. Wang, S. Makowski, A. Cieślik, H. Lv, and Z. Lv, "Fake news in virtual community, virtual society, and metaverse: A survey," *IEEE Trans. Comput. Social Syst.*, vol. 11, no. 4, pp. 4828–4842, Aug. 2024.

[87] J. Wang, S. Liu, X. Xie, and Y. Li, "Evaluating AIGC detectors on code content," 2023, *arXiv:2304.05193*.

[88] M. Khalil and E. Er, "Will ChatGPT get you caught? Rethinking of plagiarism detection," in *Proc. Int. Conf. Human-Comput. Interaction*, Cham: Springer, Jun. 2023, pp. 475–487.

[89] E. Ofgang, "What is GPTZero? The ChatGPT detection tool explained by its creator. Techlearning," 2023. [Online]. Available: https://www.techlearning.com/news/what-is-gptzero-the-chatgpt-detection-toolexplained

[90] S. Gehrmann, H. Strobelt, and A. M. Rush, "GLTR: Statistical detection and visualization of generated text," 2019, *arXiv:1906.04043*.

[91] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, vol. 1, Jun. 2019, pp. 4171–4186.

[92] A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, 2019, Art. no. 9.

[93] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2023, pp. 24950–24962.

[94] Y. Ma et al., "AI vs. human–differentiation analysis of scientific content generation," 2023, *arXiv:2301.10416*.

[95] B. Guo et al., "How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection," 2023, *arXiv:2301.07597*.

[96] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[97] J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," in *Proc. Conf. Mach. Learn.*, Jul. 2023, pp. 17061–17084.

[98] L. Mindner, T. Schlippe, and K. Schaaff, "Classification of human- and AI-generated texts: Investigating features for ChatGPT," in *Proc. Int. Conf. Artif. Intell. Educ. Technol.*, Singapore: Springer, Jun. 2023, pp. 152–170.

[99] J. Yang, S. Xiao, A. Li, G. Lan, and H. Wang, "Detecting fake images by identifying potential texture difference," *Future Gener. Comput. Syst.*, vol. 125, pp. 127–135, 2021.

[100] Y. O. Bang and S. S. Woo, "DA-FDFTNet: Dual attention fake detection fine-tuning network to detect various AI-generated fake images," 2021, *arXiv:2112.12001*.

[101] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1–11.

[102] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2387–2395.

[103] S. Suganthi et al., "Deep learning model for deep fake face recognition and detection," *PeerJ Comput. Sci.*, vol. 8, 2022, Art. no. e881.

[104] C.-C. Hsu, C.-Y. Lee, and Y.-X. Zhuang, "Learning to detect fake face images in the wild," in *Proc. 2018 Int. Symp. Comput., Consum. Control*, 2018, pp. 388–391.

[105] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, "Deep fake image detection based on pairwise learning," *Appl. Sci.*, vol. 10, no. 1, 2020, Art. no. 370.

[106] Z. Xi, W. Huang, K. Wei, W. Luo, and P. Zheng, "AI-generated image detection using a cross-attention enhanced dual-stream network," *Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Oct. 2023, pp. 1463–1470.

[107] M. F. Narvaez, I. S. Olivenza, and N. McGowan, "Painting authorship and forgery detection challenges with AI image generation algorithms: Rembrandt and 17th century dutch painters as a case study," *Int. J. Interactive Multimedia Artif. Intell.*, vol. 7, no. 7, pp. 7–13, 2022.

[108] L. Guarnera, O. Giudice, and S. Battiato, "Deepfake detection by analyzing convolutional traces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 2841–2850.

[109] S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Detecting deepfake videos from appearance and behavior," in *Proc. 2020 IEEE Int. Workshop Inf. Forensics Secur.*, 2020, pp. 1–6.
[110] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. 2018 IEEE Int. Workshop Inf. Forensics Secur.*, 2018, pp. 1–7.
[111] M. S. Rana and A. H. Sung, "DeepfakeStack: A deep ensemble-based learning technique for deepfake detection," in *Proc. 7th IEEE Int. Conf. Cyber Secur. Cloud Comput./6th IEEE Int. Conf. Edge Comput. Scalable Cloud*, 2020, pp. 70–75.
[112] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," 2018, *arXiv:1811.00656*.
[113] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5037–5047.
[114] I. Ganiyusufoglu, L. M. Ngô, N. Savov, S. Karaoglu, and T. Gevers, "Spatio-temporal features for generalized detection of deepfake videos," 2020, *arXiv:2010.11844*.
[115] A. Singh, A. S. Saimbhi, N. Singh, and M. Mittal, "Deepfake video detection: A time-distributed approach," *SN Comput. Sci.*, vol. 1, no. 4, 2020, Art. no. 212.
[116] M. A. Raza and K. M. Malik, "Multimodaltrace: Deepfake detection using audiovisual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 993–1000.
[117] T. Lavergne, T. Urvoy, and F. Yvon, "Detecting fake content with relative entropy scoring," *Pan*, vol. 8, no. 27-31, 2008, Art. no. 4.
[118] C. Vasilatos, M. Alam, T. Rahwan, Y. Zaki, and M. Maniatakos, "HowkGPT: Investigating the detection of ChatGPT-generated university student homework through context-aware perplexity analysis," 2023, *arXiv:2305.18226*.
[119] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review," *IEEE Access*, vol. 10, pp. 25494–25513, 2022.
[120] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, and Y. K. Dwivedi, "Deepfakes: Deceptions, mitigations, and opportunities," *J. Bus. Res.*, vol. 154, 2023, Art. no. 113368.
[121] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, 2018, pp. 1–6.
[122] Y. Lin et al., "Blockchain-aided secure semantic communication for AI-generated content in metaverse," *IEEE Open J. Comput. Soc.*, vol. 4, pp. 72–83, 2023.
[123] P. Radanliev, "The rise and fall of cryptocurrencies: Defining the economic and social values of blockchain technologies, assessing the opportunities, and defining the financial and cybersecurity risks of the metaverse," *Financial Innov.*, vol. 10, no. 1, 2024, Art. no. 1.
[124] S. Cai and W. Cui, "Evade ChatGPT detectors via a single space," 2023, *arXiv:2307.02599*.
[125] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," 2023, *arXiv:2307.15043*.
[126] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, "Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2023, pp. 27469–27500.
[127] W. Liang, M. Yuksekgonul, Y. Mao, E. Wu, and J. Zou, "GPT detectors are biased against non-native english writers," *Patterns*, vol. 4, no. 7, 2023.
[128] L. J. Y. Flores and Y. Hao, "An adversarial benchmark for fake news detection models," 2022, *arXiv:2201.00912*.
[129] Q. Liao et al., "Imperceptible adversarial examples for fake image detection," 2021, *arXiv:2106.01615*.
[130] S. Hussain, P. Neekhara, M. Jere, F. Koushanfar, and J. McAuley, "Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3347–3356.
[131] A. Gandhi and S. Jain, "Adversarial perturbations fool deepfake detectors," in *Proc. 2020 Int. Joint Conf. Neural Netw.*, 2020, pp. 1–8.
[132] S. A. Shahriyar and M. Wright, "Evaluating robustness of sequence-based deepfake detector models by adversarial perturbation," in *Proc. 1st Workshop Secur. Implic. Deepfakes Cheapfakes*, 2022, pp. 13–18.
[133] P. Neekhara, B. Dolhansky, J. Bitton, and C. C. Ferrer, "Adversarial threats to deepfake detection: A practical perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 923–932.
[134] N. H. Vo, K. D. Phan, A.-D. Tran, and D.-T. Dang-Nguyen, "Adversarial attacks on deepfake detectors: A practical analysis," in *Proc. Int. Conf. Multimedia Model.*, 2022, pp. 318–330.
[135] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, Aug. 2020.
[136] R. Ross, M. McEvilley, and J. Oren, "Systems security engineering: Considerations for a multidisciplinary approach in the engineering of trustworthy secure systems," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep., 2016.
[137] T. Wu et al., "Does human-AI trust affect human-AI interaction in the metaverse? Insight from a pilot study," in *Proc. 4th Int. Symp. Big Data Artif. Intell.*, 2024, pp. 204–209.
[138] D. Shin, "The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI," *Int. J. Hum.-Comput. Stud.*, vol. 146, 2021, Art. no. 102551.
[139] H. Choung, P. David, and A. Ross, "Trust in AI and its role in the acceptance of AI technologies," *Int. J. Hum.–Comput. Interaction*, vol. 39, no. 9, pp. 1727–1739, 2023.
[140] T. Zhou and H. Lu, "The effect of trust on user adoption of AI-generated content," *Electron. Library*, vol. 43, no. 1, pp. 61–76, 2025.
[141] S. Teerapittayanon, B. McDanel, and H.-T. Kung, "Distributed deep neural networks over the cloud, the edge and end devices," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst.*, 2017, pp. 328–339.
[142] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," 2015, *arXiv:1510.00149*.
[143] B. Jacob et al., "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2704–2713.
[144] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.
[145] N. Dragoni et al., "Microservices: Yesterday, today, and tomorrow," in *Present and Ulterior Software Engineering*, Cham: Springer International Publishing, pp. 195–216, 2017
[146] T. Stockhammer, "Dynamic adaptive streaming over http– standards and design principles," in *Proc. 2nd Annu. ACM Conf. Multimedia Syst.*, 2011, pp. 133–144.
[147] "Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems," 2024. Accessed: Jun. 10, 2024. [Online]. Available: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
[148] Ethics guidelines for trustworthy AI, European Commission, 2024. Accessed: Jun. 10, 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
[149] M. Alkaeed, A. Qayyum, and J. Qadir, "Privacy preservation in artificial intelligence and extended reality (AI-XR) metaverses: A survey," *J. Netw. Comput. Appl.*, vol. 231, 2024, Art. no. 103989.
[150] S. Hassan and P. De Filippi, "Decentralized autonomous organization," *Internet Policy Rev.*, vol. 10, no. 2, pp. 1–10, 2021.
[151] L. Dewitt Prat, O. N. Ndlovu Lucas, C. Golias, and M. Lewis, "Decolonizing LLMs: An ethnographic framework for AI in African contexts," in *Proc. Ethnographic Praxis Ind. Conf.*, 2024, vol. 2024, no. 1, pp. 46–85.
[152] S. D. M. T. K. Mahelona and G. Leoni, "Whisper is another case study in colonisation," 2023. Accessed: Jul. 8, 2025. [Online]. Available: https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/
[153] N. Couldry and U. A. Mejias, "The costs of connection: How data is colonizing human life and appropriating it for capitalism," in *The Costs of Connection*. Stanford, CA, USA: Stanford Univ. Press, 2019.
[154] Y. Lin et al., "Blockchain-based efficient and trustworthy AIGC services in metaverse," *IEEE Trans. Serv. Comput.*, vol. 17, no. 5, pp. 2067–2079, Sep./Oct. 2024.