Intelligent Audio-Visual: Prospects and Challenges of AIGC in Short Video Production Postprint

Authors: Zhao Yiying

Date: 2025-07-09T00:00:00+00:00

Abstract

Objective: This paper aims to explore the specific applications and challenges of AIGC technology in the short-form video domain, contemplate enhancement paths for audiovisual creation in the era of artificial intelligence, examine the risks that AIGC introduces to short-form video production, and attempt to propose technical solutions.

Methods: This paper delves into the technical operational mechanisms of AIGC, analyzing the prospects, issues, and challenges it presents for short-form video creation in conjunction with specific practices.

Results/Conclusion: The author contends that the introduction of AIGC will unlock the potential of short-form video creation in creative ideation, visual evaluation, character development, scene depiction, virtual-real integration, visual restoration, and human-computer interaction, while simultaneously confronting challenges posed by model bias, deepfakes, and digital infringement.

Full Text

Intelligent Audio-Visual: Application Prospects and Challenges of AIGC in Short Video Production

(Southwest University, Chongqing 400715)

Abstract

[**Objective**] This paper aims to explore the specific applications and challenges of AIGC technology in the short video domain, examine pathways for enhancing audio-visual creation in the artificial intelligence era, scrutinize the risks that AIGC brings to short video production, and propose potential technical

solutions. [**Method**] This paper delves into the operational mechanisms of AIGC technology and analyzes its prospects, problems, and challenges for short video creation through concrete practice. [**Result/Conclusion**] The author argues that the introduction of AIGC will unlock potential in short video creation across creative discovery, visual evaluation, character development, scene depiction, virtual-real integration, visual restoration, and human-computer interaction, while simultaneously facing challenges from model bias, deepfakes, and digital infringement.

Keywords: Generative AI; Short Video; AIGC; Large Model
Classification Code: G202
Document Code: A
Article ID: 1671-0134(2025)03-137-05
DOI: 10.19483/j.cnki.11-4653/n.2025.03.030
Citation Format: Zhao Yiying. Intelligent Audio-Visual: Application Prospects and Challenges of AIGC in Short Video Production [J]. China Media Technology, 2025, 32(3): 137-140, 158.

1. Key Technological Breakthroughs of AIGC in the Audio-Visual Domain

AIGC (Artificial Intelligence Generated Content) refers to technology that generates creative text, images, audio, video, and other multimodal AI products based on massive data, algorithms, and models [**?**]. The emergence of AIGC in 2024 is closely linked to continuous breakthroughs in its technical capabilities, which have gradually demonstrated diverse application potential. From the initial AI writing achieving "text-to-text," to AI drawing realizing "text-to-image," then to AI audio achieving "text-to-sound," and finally to the current "text-to-video" in the AI video domain, this series of developments follows a progressive technical evolution path from shallow to deep. This section will focus on introducing the key technological breakthroughs of AIGC in the audio-visual domain.

In February 2024, OpenAI released Sora, the first text-to-video model, along with 48 text-to-video cases, achieving industry-spanning breakthroughs in quality and duration, marking a sectoral leap for artificial intelligence technology in the "Text-to-Video" domain [**?**]. This has drawn increasing global attention from the media industry to the potential of AIGC technology in audio-visual content creation. Meanwhile, "text-to-video" models have also exposed certain technical weaknesses: first, the challenge of maintaining consistency in moving shots and long takes; second, spatial logic errors in generated content; and third, limitations in computational resources and production efficiency [**?**]. Consequently, compared to industrial-level audio-visual applications, AIGC is more favored in the short video creation field. For example, Kuaishou' s "Keling" text-to-video product attracted over 3.6 million users within six months of its release, gen-

erating 37 million videos [?], fully demonstrating AIGC' s enormous potential and market value in the short video domain.

Examining more deeply, as artificial intelligence technology permeates the short video field extensively, the issue of mutual adaptation between technology and society will become increasingly prominent and cannot be ignored. For this reason, this paper will delve into the technical fabric of artificial intelligence, analyzing the application prospects of AIGC technology in short video audio-visual creation while interrogating the socio-technical risks and ethical concerns it raises, attempting to provide beneficial reflections for building a harmonious and sustainable human-machine-society symbiotic relationship in the audio-visual domain.

1.1 Generative Adversarial Networks (GANs) Enhance Audio-Visual Content Authenticity

Generative Adversarial Networks (GANs) represent a core architecture in AI audio-visual applications, comprising two fundamental components: a Generator and a Discriminator [?]. The Generator' s task is to learn and simulate the distribution of real-world audio-visual data through deep learning models to produce high-quality, realistic content. The Discriminator' s task is to analyze video details such as texture, color, and motion coherence to determine authenticity. During this process, the Discriminator' s evaluation criteria continuously improve, forcing the Generator to optimize its generation strategies and thereby enhancing the realism of generated content. The "adversarial" process in GANs involves finding equilibrium between the Generator and Discriminator, where the Generator attempts to "deceive" the Discriminator with realistic content while the Discriminator evaluates authenticity to provide optimization strategies.

1.2 Variational Autoencoders (VAEs) Enhance Audio-Visual Content Coherence

Variational Autoencoders (VAEs) are an advanced method specialized in optimizing audio-visual content generation. By constructing a probabilistic latent model of audio-visual data, VAEs enable the model to skillfully balance diversity and coherence during generation, creating both rich and fluid audio-visual experiences [?]. The VAE architecture contains two core components: an encoder and a decoder. The encoder compresses video content into vector representations in latent space, while the decoder precisely reconstructs audio-visual content from these latent vectors. The essence of this technology lies in using gradient optimization between reconstruction error and KL divergence to ensure that audio-visual sequences in latent space are not only diverse but also smooth and natural. Simultaneously, by finely adjusting the dimensionality and distribution of audio-visual data in latent space, the quality and efficiency of audio-visual content generation can be further enhanced.

1.3 Sequence-to-Sequence (Seq2Seq) Models Enhance Audio-Visual Content Consistency

Sequence-to-Sequence (Seq2Seq) models address temporal issues in audio-visual content generation through Recurrent Neural Network (RNN) or Long Short-Term Memory (LSTM) architectures, demonstrating efficient management of temporal dependencies between video frames and audio milliseconds. This model' s uniqueness lies in its ability to receive a series of inputs (such as previous video frames) and predict the next series of outputs (such as future video frames), thereby ensuring visual and narrative continuity and consistency in generated videos. For example, when generating specific action scenes, Seq2Seq models can ensure reasonable action sequencing and logical plot development coherence. Additionally, the training process of Seq2Seq models allows for the incorporation of conditional encoding, such as indicators for emotional changes or scene transitions, further enhancing the model' s mastery of complex narrative structures. This makes Seq2Seq models an extremely valuable tool for audio-visual projects requiring high-level temporal control and narrative depth.

1.4 Conditional Neural Networks (CNNs) Enhance Audio-Visual Content Uniqueness

Conditional Neural Networks integrate additional conditional information such as user preferences, scene descriptions, or specific topic labels into the video generation process, greatly improving content customization and precise control capabilities [**?**]. This technology first involves encoding specific conditional information, such as vectorizing particular text descriptions or feature extraction to make them suitable for neural network processing. The encoded conditional information is integrated into the input or multiple layers of the audio-visual generation network, allowing the network to consider these guiding factors throughout the generation process. For example, when generating videos about "natural landscapes," the network incorporates relevant images and scenes tagged with "nature," or when producing videos for specific cultural festivals, it adds symbols and elements of that culture. This approach not only optimizes video visual appeal and content relevance but also enables videos to precisely meet specific scene requirements, thereby enhancing content adaptability and personalized experience.

2. Application Prospects of AIGC in Short Video Production

2.1 Inspiration Activation: Creative Discovery and Visual Evaluation

Creativity is the lifeblood of short videos, directly determining whether they can attract audiences and stand out in the competitive content market. In traditional content creation workflows, creative topics typically originate from producers' personal experiences and observations of social reality, a method that relies on creators' sensitivity and experience but is limited by individual perspec-

tives and cognitive scope, potentially leading to insufficient content diversity and innovation [**?**]. The introduction of AIGC technology will profoundly transform the creative discovery and topic planning process for short videos. This change is first reflected in the diversification of creative sources. AI algorithms can analyze social media trending topics, search engine events, user online behavior patterns, and other data to identify novel topic perspectives overlooked by the market, generating content ideas with broader coverage and fresher angles. Second, AIGC technology can also enable more efficient decision-making support in creative selection. Through machine learning models, AI systems can predict the potential audience size and popularity of different creative topics, helping producers quickly and accurately make more market-potential choices among numerous possible options.

After creativity is determined, the key lies in how to effectively transform these ideas into visual content to form short video products. Visual evaluation plays a critical role in this process. With AIGC' s assistance, creators can use natural language instruction fine-tuning to simulate different visual expression schemes, obtaining different video content alternatives at low cost and high efficiency, and optimizing through comparison. This not only helps creators conduct rapid experimentation and selection among different creative options, greatly shortening the time from concept to product, but also ensures that the final audio-visual content can precisely express creative intentions and possess market appeal.

2.2 Character Development and Scene Depiction

In short video production, character and scene design are core elements for capturing audience attention. The integration of AIGC technology injects unprecedented personalized color and infinite creative space into character development and scene depiction, opening up new artistic expression pathways for creators.

In character generation, video production teams can use AIGC technology to automatically generate virtual characters with unique personalities and appearances. This process is typically implemented based on deep learning generative adversarial networks, where the "Generator" gradually fits realistic virtual images through training on massive character data (such as specific skin tones, genders, ethnicities, etc.), while the "Discriminator" continuously feeds back evaluation results and parameter adjustments to the character image construction until training produces video characters that are both logical and full of personality. AIGC can also endow characters with more distinctive personality traits and behavior patterns based on the theme and context provided by creators. For instance, in a series of short videos showcasing different ethnic cultures, AIGC can generate virtual characters that conform to ethnic cultural characteristics and possess both "form and spirit" based on just a few keyword prompts, enhancing content authenticity and perceptibility and resonating with audiences.

Complementing character generation is scene creation and expression. AIGC

technology can automatically generate complex and delicate three-dimensional scenes according to script requirements. Using similar technologies, production teams can preview and select the most suitable backgrounds and settings for video content without physically building physical scenes, greatly reducing short video creation's dependence on external environments. Simultaneously, dynamic scene generation for specific times or atmospheres plays an important role in building video emotional appeal. Whether it' s the changing landscapes of four seasons, shuttling traffic and crowds, or rapidly changing weather phenomena, they can all be precisely captured and integrated into videos, not only increasing emotional tension but also creating an immersive experience for audiences.

2.3 Virtual-Real Integration: Bidirectional Empowerment Expands Real Boundaries

The introduction of AIGC technology significantly reduces production costs and brings unprecedented creative freedom and efficiency to short video creation. However, videos generated purely through AIGC technology are often criticized for their lack of authenticity. At this point, the advantages of traditional live-action shooting become apparent, as they can compensate for AI technology' s shortcomings in training data, optimize technical parameter feedback, and inject more vitality and vividness into short videos.

On one hand, AIGC technology demonstrates enormous potential in enhancing live-action footage. For example, in natural landscape shooting, AIGC algorithms can skillfully render natural landscapes like sunrise and sunset or add dynamic weather effects such as storms and lightning, making videos more life-like and accurately conveying shooting intentions and emotional atmosphere. Additionally, AIGC technology excels at fine-tuning picture color, lighting, and details to meet specific narrative needs or artistic styles. Using advanced physics-based simulation and particle system algorithms, AI technology can simulate realistic weather effects based on lighting and other environmental parameters at shooting locations, ensuring seamless integration of virtual elements with live-action content.

On the other hand, live-action shooting provides richer material and more authentic texture for AIGC content. Taking motion capture technology as an example, live-action data (such as details of complex human body movements, facial expressions, and environmental interactions) forms the foundation of motion models in AIGC technology [**?**]. By analyzing real human movements, AIGC technology can learn how to simulate natural motion laws, making virtual character movements not only visually natural but also physically consistent with the real world, thereby feeding back into virtual character motion generation and enhancing the authenticity and appeal of short video production.

2.4 Value Reconstruction: Visual Restoration and Content Reuse

The reuse of video materials can not only inspire new creative perspectives and reconstruct narrative structures but also significantly improve the efficiency and flexibility of short video creation. Many early short video materials, despite their extremely high content value, often suffer from poor image quality due to technical limitations, preventing these precious materials from fully realizing their potential [**?**].

Using advanced spatial enhancement technology, AIGC can apply semantic feature-based video super-resolution methods to transform originally low-resolution video materials into higher-definition, more detailed versions. This technology not only greatly improves image clarity but also preserves the unique style and flavor of the original video while making the restored video appear more vivid and realistic. In addition to spatial enhancement technology, temporal enhancement technology is another highlight of AIGC in short video visual restoration. This technology focuses on smooth transitions between video sequences, ensuring that restored videos display coherent and natural picture effects during playback. Through the organic combination of AIGC optimization networks and traditional enhancement technologies, video temporal enhancement effects have been significantly improved, enabling even complex motion scenes and rapidly switching images in old videos to be accurately and effectively repaired and restored [**?**].

2.5 Interaction Enhancement: Human-Computer Interaction and Interactive Short Videos

Interactive video is a new media form that combines video with interactive elements, aiming to create a richer and more immersive viewing experience for audiences through diversified means such as enhanced sensory feedback, deepened plot participation, and broadened content exploration [**?**]. With the vigorous development of AI technology, the field of human-computer interaction is undergoing unprecedented profound transformation, and the rise of AIGC technology has opened up new possibilities for the creation and dissemination of interactive short videos.

On one hand, AIGC technology can help achieve real-time dialogue between video characters and audiences. Through advanced technologies such as deep learning and natural language processing, AIGC can endow video characters with "intelligence." These characters can recognize and understand audience voice or text input and generate corresponding responses to achieve real-time dialogue with audiences. For example, these characters can accurately recognize and deeply understand audience voice or text input, quickly generate context-appropriate replies, and thus achieve real-time, natural dialogue with audiences. In travel short videos, audiences can easily engage in conversation with characters in the video to obtain detailed destination information and personalized travel recommendations; in educational short videos, students can interact in

real-time with teachers in the video to promptly resolve learning questions and deepen knowledge understanding; in entertainment short videos, audiences can even interact intimately with virtual idols to enjoy unprecedented unique entertainment experiences. This technology not only greatly enhances video interactivity but also allows audiences to more deeply participate in short video content and enjoy immersive viewing pleasure.

On the other hand, AIGC technology can also add rich interactive elements to video content, enhancing audience participation depth. Through precise analysis of backend data or audience connection, AIGC technology can fully understand audience needs and viewing scenarios, and then dynamically insert or adjust interactive elements in video content according to preset interaction logic or real-time audience input. These interactive elements are not limited to simple click choices but encompass diversified forms such as voice recognition and response, facial recognition and expression interaction, gesture recognition and control, and personalized content adjustment based on audience emotional feedback, providing audiences with an immersive, multi-sensory interactive experience. This new interaction model not only gives audiences more fun and sense of participation during video viewing but also enables creators to fully utilize their imagination to produce more creative and attractive interactive video works.

3. Challenges and Risks

3.1 Model Bias and Deepfakes

Text-to-video models, as dynamic presentations of images, may further exacerbate cultural, gender, socioeconomic, and racial and ethnic biases in AIGC short video content [**?**]. The bias in AIGC-generated content primarily stems from two aspects. First is data-driven bias. Text-to-video models generate new content by learning from massive video data. If these training data themselves contain biases—for example, insufficient or excessive representation of certain groups—the model-generated content will likely reflect or even amplify these biases. Second, bias in algorithm design cannot be ignored. During the model design and development stage, algorithm configuration and parameter selection completely depend on engineers' personal judgment, and individual perspectives and sociocultural backgrounds of algorithm engineers will significantly influence model bias generation.

Therefore, solving bias issues in text-to-video requires addressing these two key points. First is ensuring the comprehensiveness and diversity of training data. This includes widely collecting data from diverse cultures, regions, genders, and social groups, and ensuring the data contains diverse perspectives and narrative stories. This way, the model' s learning foundation will be more balanced, and generated video content will be more comprehensive and inclusive. Second is enhancing algorithm transparency. Developers should adopt explainable AI technology to reveal the model' s decision-making basis, allow users and stakeholders to participate in algorithm fairness testing, understand how the

model works and why it produces specific outputs, and conduct fairness-based algorithm adjustments on this foundation.

Additionally, deepfake videos are increasingly becoming the focus of attention across society. With their astonishingly realistic effects, these videos can fabricate seemingly irrefutable news or historical events, thereby misleading public understanding of major social and political issues, and in severe cases may even trigger social unrest and political crises [**?**]. Developing technologies that can effectively identify deepfake videos is key to addressing this challenge. For example, using anomaly detection technology during video inspection to precisely identify flaws in video technical processing (such as distorted facial expressions, unnatural lighting inconsistent with physical phenomena). Another approach is using comparative analysis technology to conduct detailed comparisons between video or audio samples to be inspected and known authentic samples, accurately judging their authenticity by carefully searching for possible subtle differences.

3.2 Digital Infringement and Copyright Protection

In 2024, the Beijing Internet Court heard the first AI text-to-video copyright infringement case, where the creator of the "Shanhai Qijing" trailer accused the defendant of using AI to generate animations highly similar to their work. This case triggered discussions on the originality and copyright protection of AI-generated content.

The originality issue of AIGC content is a major challenge currently facing the digital copyright protection field. On one hand, AI technology can generate works similar to or even indistinguishable from human works through learning and imitation. On the other hand, the intelligent generation process often involves complex algorithms and data processing, making it particularly difficult to judge originality [**?**]. Data rights confirmation, as a foundational process for digital copyright circulation and protection, plays an important role in protecting short video digital copyrights. First, short video creators can build data rights confirmation platforms using blockchain technology to achieve clear ownership and orderly circulation of data usage and trading systems. Second, combining cutting-edge technologies such as big data, cloud computing, and intelligent algorithms can fully model potential usage scenarios, modalities, and objects of data, more accurately identifying data value and providing scientific basis for data rights confirmation and pricing. Finally, AI technology applications can also help mainstream media achieve intelligent monitoring and rights protection of digital copyrights, promptly discovering and handling infringement actions to protect the legitimate rights and interests of data owners and content creators.

As artificial intelligence technology continues to emerge, the integration of technology and new media expression forms has become an inevitable trend. The combination of AIGC technology and short video production will comprehensively upgrade content value across creative discovery, visual evaluation, char-

acter development, scene depiction, virtual-real integration, visual restoration, and human-computer interaction. Meanwhile, social issues of the AI era, including but not limited to model bias, deepfakes, and digital infringement, are gradually becoming apparent. Therefore, while promoting technological integration and development, it is also necessary to establish corresponding technical means, supporting laws and regulations, and ethical norms to ensure the rational and orderly development of AI technology in short video creation and achieve human-machine symbiosis in society.

References

[1] Weng Yujun. AIGC, How to Embed into Normalized News Production[J]. Media Review, 2023(8): 31-33.

[2] Bai Daoxin, Zhao Su. Reflections on the Enhancement Path of Film and Television Production Based on AIGC[J]. China Media Technology, 2024(8): 138-141.

[3] Xu Bo, Li Kuangyi. The Integration and Symbiosis of AIGC and Micro-Drama: Exploration of Audio-Visual Art Driven by Technology[J]. Contemporary TV, 2024(12): 22-27.

[4] Jiemian News. Kuaishou: Keling AI Users Exceed 3.6 Million, Standalone App to Launch Soon[EB/OL]. (2024-10-24)[2024-12-13]. https://www.jiemian.com/article/11876919.html.

[5] Yuan Bin. Research and Practice on Building Video Generation Models[J]. Radio and Television Network, 2024(S1): 13-17.

[6] Wang Yanwen, Lei Weimin, Zhang Wei, et al. A Survey of Video Image Reconstruction Methods Based on Generative Models[J]. Journal on Communications, 2022, 43(9): 194-208.

[7] Shao Yu. An Exploration of Research Ideas on Generative Adversarial Neural Networks[J]. Communications and Information Technology, 2024(1): 117-122.

[8] Chen Changfeng, Yuan Yuqing. Intelligent Journalism: Generative AI Becomes Infrastructure[J]. Inner Mongolia Social Sciences, 2024, 45(1): 40-48.

[9] Wu Jianmei. Virtual Digital Character Motion Capture Technology Based on Multi-Feature Fusion[J]. Journal of Heilongjiang Institute of Technology (Comprehensive Edition), 2024, 24(1): 10-15, 21.

[10] Lin Song. Key Frame Extraction and Restoration Method for Video Images Based on Computer Vision[J]. Journal of Chongqing University of Science and Technology (Natural Science Edition), 2022, 24(6): 66-68.

[11] Wang Yong, Chen Zanwei. Thoughts on the Development and Effect Innovation of Real-Time Rendering 3D Engine Technology Application[J]. Popular Literature and Art, 2021(23): 66-68.

[12] Li Yulin. Exploration of Interactive Video Applications and Future Development[J]. Radio & TV Broadcast Engineering, 2013, 40(5): 30-32.

[13] Qin Shengyun, Li Xingyi. From ChatGPT to Sora: Production Process Reshaping and Trust Crisis Response under AIGC Transformation in the Film and Television Industry[J]. Audio-Visual, 2024(11): 3-8.

[14] Wu Jing. New Applications and Alienation Risks of "Deepfake" Technology in the Media Field[J]. Media, 2023(3): 51-54.

[15] Liu Haiming, Tao Penghui. Imitation Ethics in AIGC Copyright Practice for Media Digital Content: Controversies, Boundaries, and Principles[J]. Journalism Lover, 2024(7): 63-65.

Author Bio: Zhao Yiying (2004–), female, Han ethnicity, from Yuncheng, Shanxi, undergraduate, research interests include intelligent communication, computational communication, radio and television.

(Responsible Editor: Li Yansong)

Note: Figure translations are in progress. See original paper for figures.