

Evaluating the Impact of Multimodal AIGC Tools on the Efficiency of Short Video Production

Mingda Gao^{1,2}, Wai Yie Leong²

¹Nanchong Vocational College of Culture and Tourism, No. 1 Wenlv Avenue, Langzhong City, Sichuan Province, 637400, Nanchong, Sichuan, China

²INTI International University, Persiaran Perdana BBN Putra Nilai, 71800 Nilai, Negeri Sembilan, Malaysia

*Email: gd452292944@gmail.com, waiyie@gmail.com

Abstract

The rise of AI-generated content (AIGC) is transforming the creative industry, especially in video production. With the fast-paced demands of short videos, creators increasingly rely on AI tools. However, most are single-function, like ChatGPT for text and SUNO for music. Multimodal AIGC tools integrate text, image, audio, and video generation, streamlining workflows by automating tasks such as scriptwriting, design, music production, and editing. This study explores the effectiveness of a multimodal AIGC tool developed by ByteDance called Jimeng, examining its potential to simplify workflows, reduce costs, and foster innovation in short video creation. Using experimental methods to quantify the efficiency of short video production by multimodal AIGC tools. Based on the experiment, the multimodal AIGC tools have improved the efficiency of short video production by over 100%, revealing the ability of AIGC to reshape the short video ecosystem and enhance creative possibilities.

Keywords

AIGC, multimodal, short video, process innovation

Introduction

Since TikTok was launched in 2016, online live streaming has transitioned into the short video era (Savic, 2021). Viewers have developed a preference for concise, fast-paced, and straightforward storytelling styles. By October 2018, TikTok had been downloaded over 800 million times in more than 150 countries globally, showcasing its rapid growth in popularity (Zeng et al., 2021). Unlike traditional long-form videos, short videos require lower production costs but demand higher creativity, as creators must provide engaging and impactful content within limited time frames. The success of ChatGPT has sparked widespread interest in AI-generated content (AIGC) across various industries, which are beginning to explore integrating it into their daily operations. For example, in the field of education, AI chatbots (Vasudevan A et al., 2024). In short video production, AIGC tools enable creators to generate scripts, create visual content, compose

Submission: 21 January 2025; **Acceptance:** 26 March 2025



Copyright: © 2025. All the authors listed in this paper. The distribution, reproduction, and any other usage of the content of this paper is permitted, with credit given to all the author(s) and copyright owner(s) in accordance with common academic practice. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license, as stated in the website: <https://creativecommons.org/licenses/by/4.0/>

background music, design virtual characters, and perform AI-based editing. These technologies have improved efficiency, reduced production costs, and made the creation of high-quality content possible, effectively changing the creative process.

At this stage, most AIGC tools have only a single function, such as generating text content or generating image content. This single content generation is still relatively limited for short video production, especially in terms of maintaining content consistency. Some companies have launched their own multimodal AIGC platforms, and this paper aims to explore whether multimodal AIGC tools can further enhance the performance of short video production.

AI-Generated Content (AIGC) refers to digital content automatically produced through artificial intelligence technologies (Mingda G et al., 2024). It spans various modalities, including text, images, audio, and video (Ramesh et al., 2021). The foundation of AIGC lies in cutting-edge technologies like Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), Deep Learning (DL) (LeCun et al., 2015), Natural Language Processing (NLP), and Reinforcement Learning (RL) (Radford et al., 2019; Leong, 2024c). These tools empower machines to create content resembling human creativity and logical structure, based on inputs such as parameters, textual descriptions, or visual prompts. AIGC stands out for its “low-cost, high-efficiency, and diversity” advantages.

Multimodal AI-Generated Content (AIGC) refers to the generation of content by artificial intelligence systems that integrate multiple modalities, such as text, images, audio, and video, into a cohesive output. Unlike unimodal AIGC, which focuses on a single format, multimodal AIGC leverages input from various forms of data and produces outputs that span across different types of media. For instance, a multimodal AIGC system can process a textual description and generate a matching image or video or take an image as input to create a related narrative or audio sequence (Ramesh et al., 2021). The primary strength of multimodal AIGC lies in its ability to create richer, more dynamic, and engaging content. This technology is increasingly used in creative industries like advertising, gaming, virtual reality, and digital storytelling, where combining various media types significantly enhances user experience. For example, platforms like ByteDance’s “Jimeng” (即梦) showcase the practical application of multimodal AIGC by integrating text-to-image and video generation capabilities, providing innovative tools for content creators (Jing Sun, 2024).

At present, most of the research on AIGC focuses on applying a single AIGC tool to different fields, such as applying ChatGPT to the field of education (Yiting Q et al, 2024; Leong, 2024a), or applying AI to the medical field (Ma Cristina G et al, 2024; Leong, 2025). The common problem faced by these research fields is that the relevant types of work are not highly compatible with AIGC, and AIGC will not bring about disruptive changes to them. The short video industry is different. As an industry that relies entirely on content output for survival, there is a very large demand for the quantity and quality of content. The emergence of AIGC can essentially subvert the entire industry. At the same time, the content types produced by AIGC, such as pictures, texts, videos, and music, have a high degree of overlap with the content produced by the traditional short video industry. The use of AIGC can also improve its production efficiency. At present, the research on the application of AIGC in the media field is still relatively simple, such as using AIGC to assist in the generation of news (Suhail A. K et al, 2024; Leong, 2024b). Research on multimodal AIGC tools is very scarce, and research on multimodal AIGC is very necessary.

Experimental Design

The subjects of the experiment were 30 college students majoring in film and video production with similar production levels. All students were divided into three groups labeled ABC, with 10 people in each group. Each group needs to produce the text, video, and music for a 30-second short film. The reason for choosing college students as the subjects of the experiment is that there is a certain gap between the content produced by AI and the level of professionals in the industry. As a group with some understanding of the industry but limited skills, the content produced by college students can be compared in terms of quality with AI-generated content. The experiment mainly collected two types of data. The first type of data, T , is the average time taken by each group to produce the video, measured in minutes. The second type of data is the quality score Q for content production. Evaluating the quality of content is a relatively subjective behavior, and due to the differences in creativity and imagination among students, the produced content will inevitably vary. However, there are some essential principles for video content production, and the quality here can be understood as whether it conforms to the basic rules of short video production, such as whether the text content contains grammatical or logical errors, whether the composition of the images is standardized, whether the camera movement is reasonable, and whether there are any fundamental issues. This score is mainly used to judge whether there are any basic professional errors in the content produced by AI or students. The quality score Q is scored by 5 professional teachers in the relevant field. The quality score is divided into three dimensions.

Text Quality:Text quality refers to whether the script of a short video contains grammatical errors, whether there are unreasonable transitions between shots, and whether it can accurately express the theme of the video, etc.

Video Quality:Video quality includes whether the generated video has issues with visual content errors, whether the composition of the shots conforms to composition principles, whether the camera movement is smooth, and whether the exposure is normal, etc.

Audio Quality:Audio quality includes whether the music is clear, whether the melody and rhythm are well-matched, whether the volume of the entire audio is consistent, and whether it matches the mood of the video.

Each dimension is scored out of ten, with a total score of 30 points. The quality score Q of each group is the tie value of the quality scores of all the reorganized works. The ultimate goal of the experiment is to calculate the efficiency of producing short videos in each group, and the efficiency cannot be considered solely based on the speed of production or the quality of the content. Instead, it is about how to produce better content in a shorter time. Therefore, to quantify the efficiency of each group, the concept of efficiency parameter E was designed in the experiment, which is the quality score of short video content per minute. So, we will obtain the time T required to make a video and the quality score Q of the work, and divide the quality score Q by the required time T to obtain the production efficiency E (points/min), that is

$$E = \frac{Q}{T} \quad (1)$$

The average production efficiency of each group A, B, and C is to divide the sum of the quality scores of each group by the sum of the required duration, that is

$$\bar{E}_x = \frac{\sum_{i=1}^n Q_{i,x}}{\sum_{i=1}^n T_{i,x}} \tag{2}$$

Among them, x represents the three groups A, B or C, and n represents the number of works in each group. In this experiment, each group has ten works, that is, $n=10$. Finally, based on the experimental results, we obtained relevant data (Table 1).

Table 1: Experimental data results

Group	Average Production Time T(minutes)	Average Quality Score Q(points)	Average Production Efficiency E(points/min)
Group A (Traditional)	92	25	0.272
Group B (Single Tool)	53	23	0.434
Group C (Multimodal Tool)	41	23	0.561

Data analysis

Based on the production efficiency of each group obtained from the experiment, the specific proportion of efficiency improvement of AI tools in short video production can be calculated. The proportion of performance improvement of a single AIGC tool (Group B) compared to the traditional production method (Group A) is

$$B = \frac{0.434-0.272}{0.272} \times 100\% = 59.6\% \tag{3}$$

Compared with the traditional production method (Group A), the proportion of performance of the multimodal AIGC tool (Group C) is improved by

$$C = \frac{0.561-0.272}{0.272} \times 100\% = 106.3\% \tag{4}$$

The proportion of performance improvement of multimodal AIGC tools compared with single AIGC tools is

$$C_B = \frac{0.561-0.434}{0.434} \times 100\% = 29.3\% \tag{5}$$

From this, we can see that the AIGC tool has greatly improved the efficiency of short video production, especially the application of multimodal AIGC tools, which has improved the efficiency by 29.3% based on the single AIGC tool, and better achieved cost reduction and efficiency improvement.

Results and Discussion

The data shows that AIGC tools greatly enhance the speed of short video creation, with the highest speed increase occurring in video generation. AI can directly generate videos based on text descriptions, without the need to find shooting locations or angles, nor to determine machine parameters and composition of the scene based on the lighting conditions on the spot, saving a lot of time. In terms of script generation, AIGC tools do not spend too much time, but when generating more complex video scripts, manual assistance is required for modification, which takes more time. Compared to single AIGC tools, multimodal AIGC tools can also save more time, mainly in terms of the time required to switch between tools. Overall, AIGC tools can save a considerable amount of time in short video production (Figure 2).

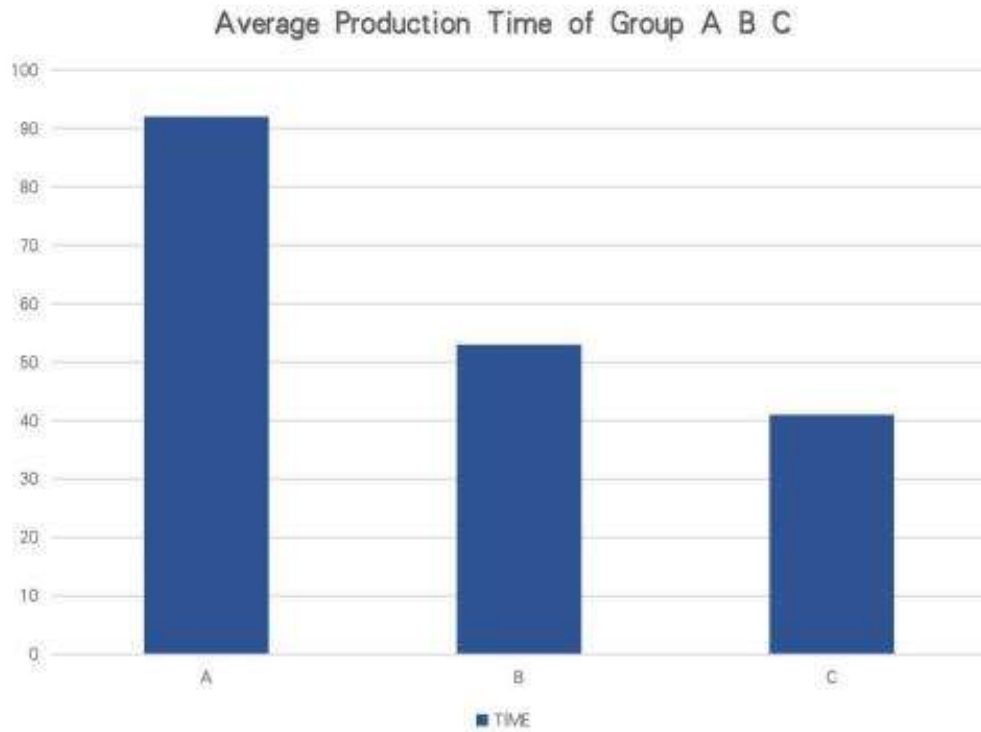


Figure 2. Average Production Time of Group A B C

In terms of quality, AIGC still has many issues. Although AI can generate scenes that are difficult to capture in real life, such as the scene of an original forest, the videos generated by AI still have many shortcomings in hair texture and character action details. This makes it easy for viewers to discern that the video was generated by AI. Especially in terms of character body details and scene depth of field, there is still a significant gap between AI-generated videos and those shot on location. Additionally, the texture of the generated videos is not realistic, and the fluidity of

motion is poor, with insufficient frame rates. On the script level, AI tends to use simpler language to explain script content, with poor continuity between shots and weak logical connection. From the experimental results, it can be seen that at the current stage, AIGC can be a good auxiliary tool for the creation of short videos, but it cannot completely replace manual production (Figure 3).

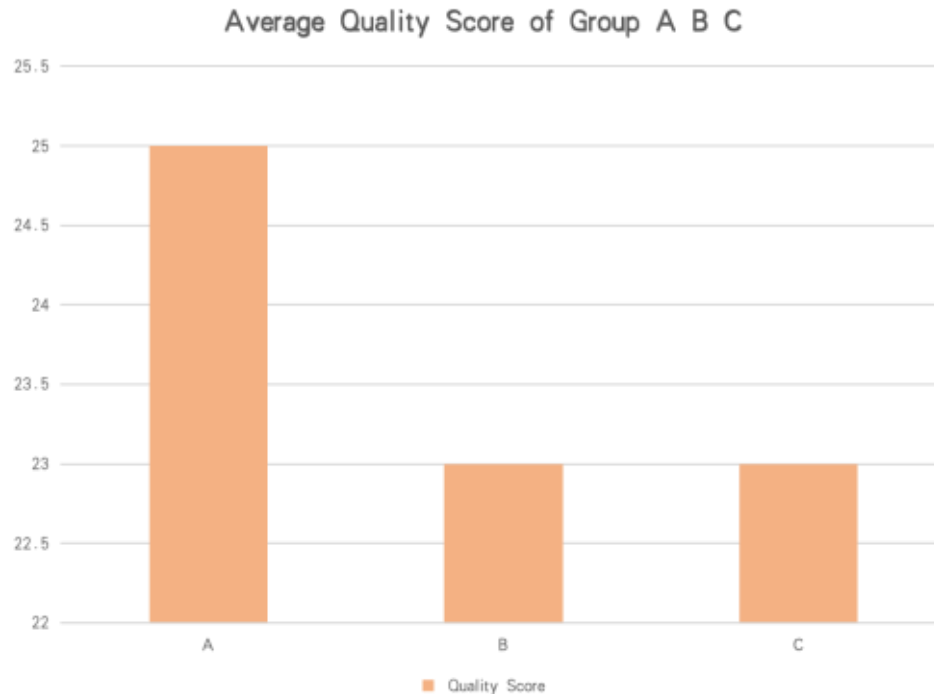


Figure 3. Average Quality Score of Group A B C

In addition to image detail issues, videos generated by AIGC also have significant problems ensuring consistency between the beginning and end of the video. Even multi-modal AIGC tools find it difficult to guarantee this. For example, it is challenging to ensure that a girl running is wearing the same clothes in different generated shots (Figure 4).



Figure 4. AI-generated videos have inconsistent characters in different shots.

At the sound level, AIGC can well complete its work. Most traditional short video professionals would choose to obtain relevant music resources from the internet. The quality of these resources varies greatly, with some very excellent and others relatively average. The choice of music is almost entirely dependent on the producer's personal musical aesthetics, perception, and judgment of musical emotions. This judgment varies from person to person, with some catering to the general taste and others being more unique, which also leads to different producers possibly choosing completely different music for the same video. Although the music generated by AIGC is slightly simplistic and popular in arrangement, it can better match the emotional expression of the video and to some extent avoid the copyright risks that existed in the past.

Conclusion

Based on the experiment, we can conclude that traditional methods maintain a relatively high production quality but take more time, leading to lower overall efficiency. Individual AIGC tools can significantly shorten production time, but they are relatively complex to operate and have certain functional limitations. Multimodal AIGC tools offer the highest speed improvement, but there is still a quality gap compared to traditional production methods. Currently, multimodal AIGC tools are more suitable as auxiliary tools for the creation of short videos. At the same time, considering the time saved, certain types of short videos are more suitable for use with multimodal AIGC tools, such as advertising-style information stream short videos. Future research on multimodal AIGC tools can consider aspects such as improving the quality of video generation, the accuracy of textures and movements, and ensuring consistency of content across multiple shots. In the future, multimodal tools will undoubtedly be a highly competitive choice for short video production.

In summary, although multimodal AIGC tools cannot completely replace human creativity, their value as auxiliary tools is undeniable. The hybrid approach of combining the efficiency of AIGC tools with the subtle creativity of human input represents a more effective strategy in the short video industry. With technological advancement, the collaboration between humans and AI will continue to redefine the boundaries of creativity and efficiency in the digital age.

References

- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2672-2680. <https://doi.org/10.48550/arXiv.1406.2661>
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. <https://doi.org/10.1038/nature14539>
- Leong, W.Y., Leong, Y.Z., Leong, W.S. (2024a). Unveiling the Intelligence Mechanisms Behind Optical Illusions, 2024 IET International Conference on Engineering Technologies and Applications, Taipei, Taiwan, October 25-27, 2024. <https://doi.org/10.1049/icp.2024.4137>

- Leong, W.Y., Leong, Y.Z., Leong, W.S. (2024b). AI in Optical Illusion Creation”, 7th International Conference on Knowledge Innovation and Invention 2024 (ICKII 2024), Nagoya Japan, AUG 16-18, 2024.
- Leong, W.Y., Zhang, J.B. (2025). Ethical Design of AI for Education and Learning Systems, ASM Sc. J., Vol. 20, No. 1, 2025. <https://doi.org/10.32802/asmscj.2025.1917>
- Leong, W.Y. (2024c). Industry 5.0: Design, standards, techniques and applications for manufacturing. Institution of Engineering and Technology, 2024. <https://doi.org/10.1049/PBME026E>
- Ma Cristina, G., See, J., Kuan, T. S., Li, M., & Lin, C. (2024). Application of Artificial Intelligence in Healthcare Industry: A Critical Review. Journal of Business and Social Sciences, 2024(35), 1-9. <https://doi.org/10.61453/jobss.v2024no35>
- Mingda, G., & Wai Yie, L. (2024). Research on the Application of AIGC in the Film Industry. Journal of Innovation and Technology, 2024(22), 1-12 <https://doi.org/10.61453/joit.v2024no22>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 9. <https://api.semanticscholar.org/CorpusID:160025533>
- Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021). Zero-shot text-to-image generation. arXiv preprint arXiv:2102.12092. <https://doi.org/10.48550/arXiv.2102.12092>
- Savic, M. (2021). From Musical.ly to TikTok: Social Construction of 2020's Most Downloaded Short-Video App. International journal of communication (Online), 3173-3195. <https://api.semanticscholar.org/CorpusID:236932324>
- Suhail, A. K., Chitra, K. E., & Wan Nor Al-Ashekin, W. H. (2024). Voice-Assisted News App using Natural Language Processing. INTI JOURNAL, 2024(24), 1-7. <https://doi.org/10.61453/INTIj.202424>
- Sun, J. (2024). AIGC Fusion Exploration: The Intersecting Path of Digital Humanities and Artificial Intelligence. J. Electrical Systems, 20(2), 327-335. <https://doi.org/10.52783/jes.1181>
- Vasudevan, A., Lama, A. V., & Sain, Z. H. (2024). The Game-Changing Impact of AI Chatbots on Education ChatGPT and Beyond. Journal of Information Systems and Technology Research, 3(1), 38-44. <https://doi.org/10.55537/jistr.v3i1.770>
- West, D. M. (2023). The AI revolution in media and entertainment: How AIGC is reshaping the industry. Brookings Institute Reports.

Yiting, Q., Munir Hayet, K., Shuqing, Z., SiYuan, C., & Choonkit, C. (2024). Enhancing Sustainability in Academic Guidance: Develop an AI-Driven Agent for Education 5.0. INTI JOURNAL, 2024(40), 1-10. <https://doi.org/10.61453/INTIj.202440>

Zeng, J., Abidin, C., & Schäfer, M. S. (2021). Research perspectives on TikTok & its legacy apps| research perspectives on TikTok and its legacy apps–introduction. International Journal of Communication, 15, 12. : <https://doi.org/10.5167/uzh-205427>