Journal of Sustainable Online Learning and Educational Research
Volume 1, Issue 1, 23-32
e_ISSN: xxxx-xxxx
https://ejournal.anjasana.com/index.php/jsoler/index

Generative AI-Based Multimedia Content Creation Automation System Development Framework

Annisa Rohiimah Sufriyani*
Universitas Bengkulu
INDONESIA

Athiya Rahma Aulia
Universitas Bengkulu
INDONESIA

Atiqah Najla Fadhilah Rachman
Universitas Bengkulu
INDONESIA

Royana Dwi Rohmah
Universitas Bengkulu
INDONESIA

Ghina Salwa Salsabilla
Universitas Bengkulu
INDONESIA

Mohammad Qais Rezvani
Rana University
AFGHANISTAN

Article Info

Article history:

Received: January 29, 2026
Revised: February 14, 2026
Accepted: February 27, 2026

Keywords:

AI
Digital
Framework
Generative
Multimedia

Abstract

Background: The world's demand for multimedia content is growing rapidly as digital platforms advance. However, the traditional content creation process, which includes ideas, writing, design, and post-production, still tends to be slow, expensive, and difficult to scale. While Generative AI has provided the ability to automatically create text, images, audio, and video, its fragmented use leads to fragmentation of the work process and does not provide a unified quality consistency mechanism.

Aims: This research aims to analyze the progress of Generative AI as well as design a systematic framework that can combine multi-modal models in one automated process for the production of multimedia content as a whole. This framework aims to improve the effectiveness, consistency, and scalability of content creation.

Methods: This study adopts a comprehensive literature review approach to 20 indexed scientific articles (2022–2025) that discuss generative models, multimodal large language models, workflow automation, model evaluation, and integration between modalities. Literature analysis was carried out through thematic synthesis to identify key trends, research gaps, integration challenges, as well as the need for generative model orchestration systems.

Results: The results show that although Generative AI models are undergoing rapid development, including diffusion models, multimodal LLMs, and knowledge-enhanced systems, there is no comprehensive framework that governs task chaining, quality control, and brand consistency in content creation.

Conclusion: This study shows that Generative AI has significant opportunities for content creation automation, but it requires structured integration through a comprehensive framework. The proposed framework offers an integrative structure for text-image-audio-video models, combining automation, style consistency, and quality control in a single workflow.

To cite this article: Sufriyani, A. R., Aulia, A. R., Rachman, A. N. F., Rohmah, R.D., Salsabilla, G. S., Rezvani, M. S. (2026). Generative AI-Based Multimedia Content Creation Automation System Development Framework. *Journal of Sustainable Online Learning and Educational Research, 1*(1), 23-32.

INTRODUCTION

The need for quality, fast, and scalable multimedia content continues to grow along with the development of digital platforms and the creative industry. The transition from manual production that requires a lot of labor to automated solutions based on artificial intelligence is a strategic necessity for organizations that are trying to maintain publication speed and brand identity consistency. Research on the ability Large Language Models (LLMs) indicate that the models have reached a level of maturity that allows them to generate text and control content contextually, so they could potentially become a key element in the content automation process (Minaee et al., 2025). In the multimodal domain, thorough research shows significant progress in Multimodal Large

*** Corresponding author:**
Annisa Rohiimah, Universitas Bengkulu, INDONESIA. ✉ rohiimahannisa@gmail.com

Language Models (MLLMs) that are able to integrate the understanding of text, images, audio, and video, these capabilities are a crucial foundation for developing a unified multimedia content production pipeline. However, these studies also emphasize the need for standardization of evaluations and inter-model orchestration methods (Yin et al., 2024; Yuan et al., 2026). In addition, structured knowledge incorporation such as Knowledge Graphs can strengthen the semantic grounding and reliability of multimodal outputs, especially when content must follow domain boundaries or brand guidelines, but, research forums show that this incorporation is still in its early stages and requires a more developed architecture (Chen et al., 2025).

In sensitive application areas, such as synthetic video for biomedical purposes, the literature emphasizes the potential of generative models while warning of risks related to clinical validity, bias, as well as the need for strict quality control. These findings support the argument that the use of Generative AI in the production of practical content needs to be complemented by quality assurance mechanisms and domain verification (Algethami et al., 2025). Based on these results, there is a real need for a framework that not only leverages LLM and MLLM capabilities, but also coordinates the linking of tasks between models, implements quality safeguards, and integrates knowledge bases to maintain stylistic consistency and domain compliance in multimedia content creation. This lack is the main concern in the development of the framework in this study.

The global demand for multimedia content is increasing along with the expansion of digital platforms and the creative economy. Traditional content production involving the stages of ideation, writing, graphic design, and post-production is considered no longer adequate to answer the needs of real-time and large-scale content. This challenge is in line with the findings (Hasanuddin & Nurfransiska, 2026; Yuan et al., 2026), which confirms that the industry's demand for digital content is soaring much faster than the ability of human-based production. Time, cost, and reliance on experts make conventional production processes difficult to scale efficiently. Technological advancements in Generative Artificial Intelligence (Generative AI) has opened up new opportunities to respond to these challenges. Generative AI, in particular Large Language Models (LLM) and diffusion models, have demonstrated the ability to produce text, images, and sounds that are realistic and contextually relevant. (Minaee et al., 2025) Classify Generative AI as a transformative technology that allows systems to generate new content based on patterns from training data automatically. This encourages the automation of various aspects of content production, from script preparation to illustration creation. For example, visual models such as DALL· E 3 shows a significant improvement in the synthesis of images that are more accurate to textual descriptions (Yuzyk et al., 2025), while generative video systems like Phenaki prove that video creation from text input is now becoming more realistic and flexible (Lv, 2023).

In addition, the development of multimodal technology has strengthened integration between types of media. (Sengar et al., 2025) emphasizing that the multimodal model is able to combine information from text, images, audio, and video to produce a more comprehensive output. These integrative capabilities are important in modern content production that requires consistency across media and continuity of creative flows. Thus, Generative AI not only serves as an automation tool, but also a creative support that expands the production capacity of creators and organizations. However, while the various applications of Generative AI are available individually, their fragmented use poses new challenges. Users are faced with a manual process to manage prompts, outputs, and synchronization of styles or brand identities across multiple platforms. (Hasanuddin & Nurfransiska, 2026) reveals that without a structured orchestration system, the use of generative AI technology can actually increase operational complexity. Therefore, a framework is needed that is able to systematically integrate generative AI models in the production workflow, including task chaining, quality validation, and automatic output adaptation.

The research and development of Generative AI-based content automation frameworks is becoming increasingly important to realize fast, consistent, and scalable content production. The framework is expected to serve as a backbone that brings together the capabilities of multimodal models, optimizes creative processes, and provides flexibility for organizations to produce hyper-personalized and on-demand content. Thus, the next generation of digital content production systems will be more responsive to market dynamics and the increasing need for personalization in the digital era.

Although rapid advances in Generative Artificial Intelligence (Generative AI) have resulted in powerful models, there are significant gaps that hinder systematic implementation. Research has successfully established the taxonomy of Generative AI models (Minaee et al., 2025) and recognize the great potential of Foundation Models (FMs), while highlighting the risks associated with Governance and Alignment (Hasanuddin & Nurfransiska, 2026). In terms of capabilities, literature shows great achievements in content creation Single-step, such as text-to-image with an increase Alignment via DALL· E 3 (Yuzyk et al., 2025) and variable length video generation via Phenaki (Lv, 2023). In addition, the importance of integrating various data modalities has been conceptually understood (Sengar et al., 2025). However, the main gap that emerges is the lack of Architecture systematic or framework that explicitly addresses the problem of model fragmentation and workflow orchestration (Workflow Orchestration). Existing papers focus on models Individual or fusion data, but does not provide Framework that is required to autonomously string (Chaining) Output from one model to the next multi-modal model in the process of continuous content production and end-to-end. This gap is exacerbated by the absence of a centralized control mechanism (guardrail) at the system architecture level to maintain consistency of style, tone, and brand identity across the outputs produced by different models, a challenge that goes beyond the issue of single model alignment. Therefore, the urgent need lies in designing a technical and methodological framework that can bridge the capabilities of these individual models into a cohesive and controlled content production system.

The urgent need to address the Time-to-Publish and Resource Dependency Gap was a key driver of this research, given that traditional content production processes are labor-intensive, expensive, and not scalable due to their reliance on the expertise of individual specialists. While Generative AI promises a drastic increase in efficiency, its full realization is hampered by the Fragmented Workflow Gap. The sporadic implementation of standalone AI tools actually adds complexity because it requires manual management of prompts and outputs between platforms, thus thwarting the potential for automated workflows. Therefore, the rationalization of this research lies in the need to design a framework that serves as a backbone of the system to intelligently manage the chaining of multi-modal models (text, visual, audio). Furthermore, this research also overcomes the Brand Consistency and Quality Guardrail Gap. It's important to ensure that automation doesn't sacrifice quality; The framework should provide a centralized mechanism to maintain consistency of style, tone, and brand identity across outputs, a crucial function for AI adoption in the corporate environment. Thus, this study is particularly relevant to provide a comprehensive technical and methodological blueprint for building an efficient, fully automated, scalable, and brand-compliant Generative AI-based multimedia content production system, filling a critical gap in the creative automation system development literature.

This research aims to design and develop a Generative AI-based systematic framework that is able to automate the entire process of creating multimedia content end-to-end. The framework is designed to integrate a variety of multi-modal models, from text, images, audio, to video, into one integrated pipeline. This integration is expected to be able to improve production efficiency, expand system scalability, and maintain consistency in the quality of the content produced. In addition, this research also focuses on the preparation of technical and methodological blueprints that can be used by organizations and developers to build a fully automated, hyper-personalized, and brand-consistent content production system.

The first hypothesis (H1) states that the integration of Generative AI models in an automation framework will be able to significantly reduce the production time of multimedia content compared to traditional production methods. This time efficiency is rated as one of the main indicators of automation success. The second hypothesis (H2) states that a structured multi-modal framework can reduce workflow fragmentation resulting in a more efficient production flow than the use of AI tools that work standalone. The third hypothesis (H3) highlights the importance of the chaining model system in the framework, which is thought to be able to improve the consistency of style, tone, and brand identity in the content produced. This consistency is often difficult to achieve in manual or semi-automated processes. Furthermore, the fourth hypothesis (H4) suggests that the use of a generative framework approach will objectively improve the quality of multimedia content output, both in terms of contextual suitability, narrative structure, and visual aesthetics, when compared to

non-integrated pipelines. Finally, the fifth hypothesis (H5) states that Generative AI-based automated content production frameworks have the potential to scale up content production capacity. With full automation, organizations can produce high-volume content without requiring a proportionate increase in operational costs. If proven, this will make the framework a strategic solution for organizations that require large-scale and sustainable content production.

METHOD

The research method used in this study is a comprehensive literature review, which aims to identify, evaluate, and synthesize various relevant scientific findings related to the development of a Generative AI-based multimedia content creation automation framework. The selection of literature review as a research design is based on the need to understand the theoretical and practical developments of Generative AI technologies in the fields of text, images, audio, and video, as well as how the technology has been used in various multimodal architectures and automation pipelines. Through this approach, research can gather the latest insights needed to design new frameworks that are integrative and able to answer various limitations that exist in conventional content production systems and AI solutions that are still fragmented.

The literature review process is carried out through several systematic stages, starting from tracing academic sources in various scientific databases such as Scopus, SpringerLink, IEEE Xplore, Elsevier ScienceDirect, Taylor & Francis, ACM Digital Library, Wiley Online Library, as well as other scientific repositories such as Google Scholar and arXiv. The focus publication time range is year 2022 to 2025, because this period reflects the rapid development phase of modern Generative AI models, including diffusion models, transformer-based LLMs, multimodal LLMs, and video-audio generative models. Literature inclusion criteria include: (1) articles have a valid DOI, (2) the focus of the study is related to Generative AI, multimodal learning, content production automation, or model orchestration systems, (3) articles come from reputable indexed journals or conferences, and (4) research makes theoretical or practical contributions to the construction of AI-based automation frameworks. From this selection process, the research identified various key sources such as publications by (Algethami et al., 2025; Chen et al., 2025; Hasanuddin & Nurfransiska, 2026; Tang et al., 2023; Yuan et al., 2026; Zhang et al., 2022; Zhao et al., 2025) which provides a solid foundation for analysis.

The next stage is Critical Evaluation to the selected literature, which was carried out by examining the theoretical depth, research methodology, technical accuracy, and relevance of the content to the focus of the study. This evaluation includes an understanding of the working mechanisms of generative models such as GANs, diffusion models, transformer-based LLMs, and multimodal fusion frameworks. In addition, the study also assesses how various studies apply these concepts in real-world scenarios such as digital animation creation, video synthesis, image generation for the creative industry, and text-to-video pipeline. The focus of the evaluation is also directed at key issues emerging from the literature, including the challenges of workflow fragmentation, limitations of cross-modal integration, output quality and consistency issues, and the lack of style control and brand identity in generative systems. Research from (Dongoran et al., 2022) (Rahmani & Liu, 2026) as well as several multimodal surveys (Yao, 2024) (Yuan et al., 2026) becomes important in describing the practical and technical gaps that have not been solved.

The literature is then synthesized using a thematic synthesis approach, where scientific findings are grouped into several major themes such as: the development of multi-modal generative models; text image audio video integration architecture; workflow automation; stylistic consistency and content quality control; and the challenges of generative model orchestration in the production pipeline. The analysis also identifies important conceptual trends such as knowledge-enhanced multimodal learning, metadata-guided content generation, hierarchical orchestration in AI systems, and multi-agent generative pipelines that are starting to emerge as new approaches in the 2024–2025 research. From the results of this synthesis, the research found that although many studies have explored the capabilities of generative models in each modality, there is no single architectural framework that integrates all of these modalities in one consistent, adaptive, and industrially scalable automated system.

The last stage of this literature review method is the formulation of a conceptual framework as the main contribution of the research. The framework is compiled based on the research gaps found, and integrates technical insights from various related studies. In this stage, the researcher designed a framework that includes a modular structure for multi-modal model integration, an automated chaining mechanism between models, AI-based quality control management, a brand consistency system, as well as an automated content publication pipeline. The design of the framework resulted from an in-depth analysis of existing theories, technologies, and practices, as well as a critical evaluation of the shortcomings of generative AI systems that have been developed sporadically. Thus, literature review not only produces a summary of previous research but also forms a valid methodological basis for designing an efficient, scalable, and futuristic Generative AI-based multimedia content production framework.

RESULTS AND DISCUSSION

Results

The twenty papers studied show rapid developments in the fields of Generative AI, multimodal learning, video generation, diffusion models, as well as educational and biomedical applications. In general, this literature can be divided into four main clusters: (1) generative models for video and images, (2) multimodal learning and multimodal large language models, (3) the application of AI in multimedia content, and (4) domain-specific applications such as biomedical and education. Each cluster makes an important contribution, but it also shows various limitations that become the next research opportunity. Papers about generative video such as by (Gou et al., 2024; Luo et al., 2025; Sun et al., 2024; Zhao et al., 2025) offers a thorough overview of the development of generative video models, especially Diffusion models. Their strength lies in the comprehensive mapping of current techniques and challenges such as Temporal consistency and computational scale. However, most of these studies are highly technical and lack the ethical, policy, and social impact aspects of increasingly realistic generative video. In addition, although they mention performance challenges, they have not provided a truly concrete research direction in terms of resource optimization and energy efficiency.

In the multimodal learning and multimodal LLM clusters, major contributions came from (Chen et al., 2025; Lymperaiou & Stamou, 2024; Tang et al., 2023; Yuan et al., 2026). These papers show a major shift from the unimodal model to the Multimodal Large Models (MM-LLMs) that are capable of processing text, images, audio, and video all at once. Critically, they are strong in describing architecture (late fusion, early fusion, co-attention, knowledge-enhanced learning), but still lacking in the discussion of standardized evaluation metrics. Another gap that arises is the lack of a truly efficient multimodal model for real-time deployment, especially on edge or mobile devices. Some papers also focus too much on large models but do not address the issue of multimodal data privacy that should receive attention in public applications. In the realm of generative AI for multimedia and creative content, papers such as (Dongoran et al., 2022; Nugroho, 2025; Tiwari & Misra, 2018) examine how AI is starting to automate the creation of images, videos, animations, and other visual content. The positive side of these studies is the explanation of the practical application of generative AI in the creative industry and the optimization of production workflows. However, some are still descriptive, not analytical, because they do not provide quantitative evaluations or data-driven case studies. Ethical challenges such as work ownership and plagiarism are also not discussed in depth even though they are a big issue in AI-generated content.

Paper in a custom application domain, such as (Algethami et al., 2025) which discusses synthetic videos in the biomedical field and (Yao, 2024) about multimodal teaching, expanding the context of the application of generative AI. Both show how AI is not only used for entertainment but also education and health. The advantage of these studies is that they provide an overview of real use cases and industry needs. However, they also point to real limitations: generative models still have a risk of bias, a lack of clinical validation (for biomedicine), as well as limitations in the adaptability of multimodal learning in highly diverse educational environments. Two other papers, such as (Zhang et al., 2022) and paper about Knowledge Graphs for (Lymperaiou & Stamou, 2024) provide a structural perspective on how generative models are integrated with structured knowledge to enrich

the understanding of the model. This approach is innovative, but it is still under-evaluated on a large scale and lacks robust experiments on robustness and explainability in multimodal contexts.

Overall, the main trends that appear throughout the literature are the dominance of diffusion models, the emergence of multimodal LLMs, and the integration of generative AI with domain-knowledge such as knowledge graphs. The weaknesses are the lack of evaluation standardization, the issue of bias and ethics that are still often ignored, and the large need for compute which is a challenge in widespread implementation. Visible research gaps include: the need for lighter generative models, ethical-in-design integration, more consistent multimodal evaluation systems, and exploration of applications that are more practical and industry-adoptable at scale. Thus, these twenty papers collectively provide a strong theoretical and technical foundation regarding the cutting-edge development of generative AI and multimodal systems. However, they also open up a wide research space related to efficiency, security, ethics, and more mature real-world applications.

Discussion

Cross-paper comparisons show fundamental differences in the technical themes discussed, particularly in the family model and the approach used. Diffusion-based papers like works (Sun et al., 2024) emphasizing diffusion's excellence in visual quality, while Shi added an innovation in the form of MM-Diffusion that handles audio-video synchronization. On the other hand, some studies such as (Nugroho, 2025), as well as some of the (Tiwari & Misra, 2018) still make room for GAN and hybrid approaches, which they consider superior for creativity and real-time potential despite being less stable than diffusion. Other groups, such as (Yuan et al., 2026), taking a different path with a focus on MM-LLM and structured pretraining, which serve as a representational backbone and multimodal controller rather than as a media sampling model. In summary, key differences emerge between research focused on the sampling process (diffusion/GAN) and research focused on multimodal representation and orchestration (MM-LLM/KG), two directions that complement each other but emphasize different research priorities. In terms of modality, video-centric papers such as (Gou et al., 2024) focuses on improving frame quality and temporal consistency. Instead, Shi extends the scope to the audio-visual domain with special attention to synchronization between modalities. Multimodal paper groups such as (Yin et al., 2024) and (Zhao et al., 2025) Examines the integration of text, images, audio, and video with an emphasis on alignment and fusion strategies. In addition, there are also domain-specific papers (Algethami et al., 2025) in biomedical and (Yao, 2024) on Education expresses the need for strict domain constraints. Level Multimodality These differences are the source of the variety of technical answers offered, multimodal paper emphasizes control and orchestration, while paper Video-centric Emphasis on visual quality.

Comparisons by application domain reveal very different goals. Paper like (Tiwari & Misra, 2018) oriented towards creative production and the animation industry, while (Algethami et al., 2025) highlighting the need for rigorous validation in sensitive biomedical domains. (Yao, 2024) On the other hand, it moves in an educational context with a focus on content adaptation. Survey studies such as (Yuan et al., 2026) Map a general landscape without a specific domain focus. These variations in domains create different standards of success, ranging from creativity, clinical accuracy, to pedagogical effectiveness. In terms of evaluation, video-centric research tends to use visual quality metrics such as FID, IS, and LPIPS, while multimodal research uses metrics such as CLIP-score, mAP retrieval, and audio-visual sync metrics as seen in (Asyhari et al., 2025; Hasanuddin & Nurfransiska, 2026) Specific paper domains such as (Algethami et al., 2025) Demand Expert-Based Validation or Human Study. However, there is no consensus on holistic metrics that can assess visual, temporal, and semantic at the same time, making it difficult to compare claims between papers. Dataset differences also play a big role: Video-centric papers use large datasets such as WebVid, UCF101, and Kinetics, while domain-specific studies such as Algethami use small, highly curated medical datasets. Multimodal papers and MM-LLMs generally rely on large multimodal datasets for pretraining, so differences in scale and data types also affect the results.

The contribution of each paper also varies between theoretical and practical. Paper like (Gou et al., 2024; Yin et al., 2024) methodological with contributions to the architecture as well as research recommendations, while (Asyhari et al., 2025; Nugroho, 2025; Tiwari & Misra, 2018) focusing on engineering practices, production pipelines, and real implementation. Domain papers such as

Algethami and Yao provide an overview of practices that are limited by regulations or pedagogical goals. The general strengths of diffusion groups and MM-LLMs are the quality of results and the ability to control multimodally, but their limitations include large computational needs, lack of comprehensive evaluation metrics, reproducibility issues, and ethical discussions that tend to be surface-based. Debates arose over the relevance of GANs to real-time needs and whether controls should be attached to MM-LLMs as orchestrators or to the generator pipeline itself.

Although different, these papers complement each other in building a multimodal automation framework. MM-LLM survey and Knowledge Graph (Chen et al., 2025) can serve as a control module and semantic grounding. Diffusion and MM-Diffusion based generators (Sun et al., 2024) provides high visual quality and synchronization. Research retrievals such as (Asyhari et al., 2025) assist in asset provisioning and indexing, while implementation papers such as Azhar and Nugroho provide guidance for production pipeline design and system integration. For sensitive domains, ethics and validation guidance from (Algethami et al., 2025; Yao, 2024) is still needed. In terms of production readiness, paper engineering such as (Asyhari et al., 2025; Dong et al., 2024; Nugroho, 2025) relatively more ready to be implemented, while computationally heavy diffusion research still requires further optimization. Domain-sensitive studies like Algethami require validation and regulation before they can be used practically. Overall, the practical implications lead to a combination of approaches: MM-LLM and knowledge graph are used as controllers, diffusion is used as a high-quality generator, retrieval for asset reuse, and paper engineering is the foundation of pipeline implementation. Evaluation must combine visual, multimodal, synchronization, and human study metrics because there is no single metric agreed. Optimizations such as distillation, quantization, and streaming on generators are needed to bridge the gap between academic quality and real-time production needs.

Implications

From the twenty papers studied, it can be seen that these studies have major implications for the direction of the development of Generative AI, especially in diffusion models, multimodal LLMs, and the integration of structured knowledge such as knowledge graphs. This trend suggests that AI systems are increasingly moving towards the ability to understand and generate complex content across modalities, potentially transforming the way media production, education, and even medical analysis are produced. However, the implications are not only technical; The research also uncovers important challenges related to ethics, data bias, multimodal privacy, and the need for evaluation standardization, which means that technology development cannot be separated from social and regulatory considerations. The emergence of large models also raises practical implications regarding computational loads, so model efficiency is a critical issue for large-scale implementation in industry and edge devices.

Research contribution

Collectively, these twenty papers make a major contribution to science through mapping the latest developments in generative models, deepening multimodal learning architectures, and integrating AI with specific domains such as education and biomedicine. Papers on diffusion models strengthen the theoretical foundations of generative sampling and temporal consistency, while studies of MM-LLMs encourage a new understanding of how large models can simultaneously manage text–visual–audio–video inputs. Another contribution came from research that addresses pipeline engineering and the practice of using AI in media production, which provides a bridge between academic innovation and industry needs. On the other hand, domain-based research such as medical and education contributes insights into the safe, relevant, and high-value application of AI to society. Overall, this collection of papers forms an ecosystem of complementary contributions between theory, model architecture, evaluation, and practical applications.

Limitations

Despite offering broad insights, these twenty papers still have some significant limitations. Many generative studies, especially those based on diffusion, tend to be highly technical and lack the ethical, privacy, and social impact aspects of using synthetic content. Multimodal papers often do not provide consistent evaluation standards and still focus on large models that are difficult to implement

in real-time or on devices with limited resources. Some studies are also descriptive without strong experimental support or direct quantitative comparisons. In specialized domains such as biomedicine, its limitations arise in the form of a lack of clinical validation and a high risk of bias. In addition, research on knowledge graphs and integrated structured knowledge has not been evaluated on a large scale, so the robustness and explainability of the model have not been fully tested. In general, the big gaps seen are the lack of standardization of metrics, the limitations of specific datasets, and the still very high computational needs.

Suggestions

Based on existing limitations, several research suggestions can be submitted for the development of generative AI and multimodal systems in the future. Further research needs to develop a more efficient and lightweight model that allows deployment on edge, mobile, or in industrial settings that prioritize production speed. There is also a need for standardization of evaluation metrics that are able to assess visual, temporal, semantic, and multimodal aspects in an integrated manner. Ethical, security, and privacy integration should be placed at the model design stage, not just a post-development add-on. For domain-specific research, empirical validation, clinical validation in biomedicine or long-term pedagogical studies in Education, is a must for technology to be adopted responsibly. In addition, the exploration of generative AI integration with knowledge graphs, retrieval systems, and MM-LLMs needs to be expanded through large-scale experiments to strengthen the robustness and reliability of the model. With these measures, future research can bridge the gap between academic performance and sustainable industry implementation.

CONCLUSION

This research shows that the main challenge in the production of modern multimedia content lies in the inefficiencies of traditional processes and the fragmentation of the use of Generative AI tools separately. Based on a review of twenty related studies, it was found that although Generative AI has made significant advances in text, images, audio, and video, there is no unified framework that can orchestrate the entire production process automatically and consistently. In line with the objectives of the study, the analysis in the Results and Discussion section succeeded in identifying gaps related to multimodal integration, the need for evaluation standardization, and the lack of quality control mechanisms and brand consistency. Based on these findings, this study formulated a conceptual framework that is able to integrate generative models in one automated workflow based on task chaining and quality guardrails. The framework offers a comprehensive solution that can improve the efficiency, scalability, and quality of multimedia content production. Future development prospects include prototyping of this framework-based system, performance testing on various production scenarios, computational optimization for real-time use, as well as expanded integration with knowledge graphs and retrieval systems to improve the accuracy and relevance of content. In addition, advanced research needs to pay attention to the aspects of ethics, multimodal privacy, and standardization of metrics, so that the framework can be implemented widely and responsibly in various sectors such as the creative industries, education, and public services.

ACKNOWLEDGMENT

The author would like to thank the Informatics Study Program of the University of Bengkulu for providing academic support and facilities during the process of preparing this research. Thank you also to the lecturer in charge of the Introduction to Multimedia Systems course, who has provided direction, feedback, and guidance during the completion of this paper. In addition, the author is grateful to all group colleagues who contributed to the discussion, literature collection, and preparation of analysis so that this research could be completed properly.

AUTHOR CONTRIBUTION STATEMENT

AR, AA, AN, RR, and GS contributed to the preparation of this study. All authors are jointly involved in determining the topic and focus of the study, collecting and reviewing the literature, thematic analysis, and formulation of writing structure. The authors also participate in the preparation of manuscripts, the integration of research findings, the consistency of the content, and the alignment of writing styles. In addition, each author provides input on the results and discussions, and plays an active role in the revision process until the final manuscript is mutually agreed.

REFERENCES

Algethami, N., Iqbal, T., & Ullah, I. (2025). Generative AI for biomedical video synthesis : a review. *Artificial Intelligence Review*, *58*(392), 1–50. https://doi.org/10.1007/s10462-025-11394-5

Asyhari, M. F., Dimas, F., Bakar, A. M. A., & Bastian, A. (2025). PENCARIAN CERDAS ANTAR-MODA : EVOLUSI TEKNOLOGI VIDEO-TEXT RETRIEVAL. *Jurnal Informatika Dan Teknik Elektro Terapan*, *13*(3), 69–79. https://doi.org/10.23960/jitet.v13i3.6607

Chen, Z., Zhang, Y., Fang, Y., Geng, Y., Guo, L., Chen, J., Liu, X., Pan, J. Z., Zhang, N., Chen, H., & Zhang, W. (2025). Knowledge Graphs for Multi-modal Learning: Survey and Perspective. *Information Fusion*, *121*, 103124. https://doi.org/10.1016/j.inffus.2025.103124

Dong, A., Wang, L., Liu, J., Lv, G., Zhao, G., & Cheng, J. (2024). MFIFusion: An infrared and visible image enhanced fusion network based on multi-level feature injection. *Pattern Recognition*, *152*, 110445. https://doi.org/10.1016/j.patcog.2024.110445

Dongoran, I. M., Azhar, I. N., Anto, J., & Hakim, D. L. (2022). The Effect of Interactive Multimedia on Student Behavior Against Covid-19 in Vocational High Schools. *Education and Humanities Research*, *651*(Icieve 2021), 130–133. https://doi.org/10.2991/assehr.k.220305.027

Gou, J., Xie, N., Liu, J., Yu, B., Ou, W., G, Z. Y., & Chen, W. (2024). Hierarchical graph augmented stacked autoencoders for multi-view representation learning. *Information Fusion*, *102*, 102068. https://doi.org/10.1016/j.inffus.2023.102068

Hasanuddin, M., & Nurfransiska, F. (2026). Pengaturan Artificial Intelligence (AI) Dalam Perspektif Hukum Indonesia : Analisis Normatif Atas Tntangan , Implikasi , Dan Model Regulasi Ideal. *Judge : Jurnal Hukum*, *06*(06), 1890–1897. https://doi.org/10.54209/judge.v6i06.1632

Luo, Y., Chen, E., & Yang, S.-H. (2025). Generative AI in Engineering Education: A Survey of Student and Instructor Usage and Attitudes. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *69*(1), 272–277. https://doi.org/10.1177/10711813251358792

Lv, Z. (2023). Generative artificial intelligence in the metaverse era. *KeAi Communications*, *3*(May), 208–217. https://doi.org/10.1016/j.cogr.2023.06.001

Lymperaiou, M., & Stamou, G. (2024). A survey on knowledge - enhanced multimodal learning. In *Artificial Intelligence Review* (Vol. 57, Issue 10). Springer Netherlands. https://doi.org/10.1007/s10462-024-10825-z

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2025). Large Language Models : A Survey. *ArXiv*. https://doi.org/10.48550/arXiv.2402.06196

Nugroho, A. Y. (2025). Deep Learning Algorithms in the Development of Generative AI Models for Automated Content Creation. *Mutiara : Jurnal Penelitian Dan Karya Ilmiah*, *3*(5), 111–122. https://doi.org/10.59059/mutiara.v3i5.2804

Rahmani, H. A., & Liu, J. U. N. (2026). AI-Generated Content (AIGC) for Various Data Modalities : A Survey. *ACM Computing Surveys*, *57*(9), 1–67. https://doi.org/10.1145/3728633

Sengar, S. S., Bin, A., Sanjay, H., & Fiona, K. (2025). Generative artificial intelligence : a systematic review and applications. *Multimedia Tools and Applications*, *84*, 23661–23700. https://doi.org/10.1007/s11042-024-20016-1

Sun, L., Lian, Z., Liu, B., & Tao, J. (2024). HiCMAE: Hierarchical Contrastive Masked Autoencoder for self-supervised Audio-Visual Emotion Recognition. *Information Fusion*, *108*, 102382. https://doi.org/10.1016/j.inffus.2024.102382

Tang, X., Rohaida, S., Mohamed, B., & Li, Q. (2023). Multimedia use and its impact on the effectiveness of educators: a technology acceptance model perspective. *Humanities and Social Sciences Communications*, *10*(1), 923. https://doi.org/10.1057/s41599-023-02458-4

Tiwari, A., & Misra, M. (2018). Analysis of operative factors and practices in social CRM. *International Journal of Digital Enterprise Technology*, *1*(1), 135–176. https://doi.org/10.1504/IJDET.2018.092639

Yao, X. (2024). Research on Multimodal English Teaching Methods and Practices Leading to Intelligent Generation. *Applied Mathematics and Nonlinear Sciences*, *9*(1), 1–18. https://doi.org/10.2478/amns-2024-1654

Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., & Chen, E. (2024). A survey on multimodal large language models. *National Science Review*, *11*(12), n403. https://doi.org/10.1093/nsr/nwae403

Yuan, Y., Li, Z., & Zhao, B. I. N. (2026). A Survey of Multimodal Learning : Methods , Applications , and Future A Survey of Multimodal Learning : Methods , Applications ,. *ACM Computing Surveys*, *57*(7), 1–34. https://doi.org/10.1145/3713070

Yuzyk, O., Honcharuk, V., Pelekh, Y., Bilanych, L., Sirenko, P., Voitovych, I., Roienk, L., Bilanych, H., Makukh, D., Zidens, J., & Yuzyk, M. (2025). Research on Generative Artificial Intelligence Technologies in Education: Opportunities, Challenges, and Ethical Aspects. *BRAIN. Broad Research in Artificial Intelligence and Neuroscience*, *16*(1), 139–151. https://doi.org/10.70594/brain/16.S1/12

Zhang, Z., Li, Z., Wei, K., Pan, S., & Deng, C. (2022). A survey on multimodal-guided visual content synthesis. *Neurocomputing*, *497*, 110–128. https://doi.org/10.1016/j.neucom.2022.04.126

Zhao, M., Wang, W., Zhang, R., Jia, H., & Chen, Q. (2025). TIA2V: Video generation conditioned on triple modalities of text–image–audio. *Expert Systems with Applications*, *268*, 126278. https://doi.org/10.1016/j.eswa.2024.126278