Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

Overview of Multimodal Generative Models in Natural Language Processing and Computer Vision

LI, Liang [1*]

[1] Shandong Youth University of Political Science, China

**** LI, Liang is the corresponding author, E-mail: liliang150851@163.com***

Abstract: Multimodal generative models have become essential in the deep learning renaissance, as they provide unparalleled flexibility over a diverse context of applications within Natural Language Processing (NLP) and Computer Vision (CV). In this paper, we systematically review the basic concepts and technical improvements in multimodal generative models by discussing their applications across different modalities such as text, images, audio,and video. These models though augment the strength of AI to comprehend and perform complicated tasks by coalescing data from various modalities. In this paper, we investigate how these principles apply to many of the existing mainstream models (including CLIP, DALL·E, Flamingo), and consider their applications in VQA,text-to-image-synthesis; medical image analysis; edutainment content creation & user research developments. This paper also examines the existing difficulties of such technologies including paucity in data availability, modality fusion effectiveness and constraints on computational resources while suggesting pathways for future research. The paper goes on to state privacy parallels between multi-modal generative models (GGMs) calls for a model of safety over responsibility when it comes to technological innovation.

Keywords: Multimodal Generative Models, Natural Language Processing, Computer Vision, Data Fusion, Deep Learning, CLIP, DALL·E.

Disciplines: Artificial Intelligence.

Subjects: Multimodal Generative Models.

DOI: https://doi.org/10.5281/zenodo.13988327

ARK: https://n2t.net/ark:/40704/JCTAM.v1n4a09

1 INTRODUCTION

In recent conferences on NLP (like EMNLP/ACL) and CV, multimodal generative models have become one of the most cutting-edge research directions, and their significance is increasing by every passing day[1]. Deep learning and artificial intelligence technology are developing rapidly, also leading to more diverse data sources which the system can work with in a better cross-modal fusion [2]. Generative models help machines to have a better understanding as they can be trained on multimodal data, such as text, images and audio using video in one model which is giving more contextual information enabling them to generate output that are very natural; semantically profound [3]. The multimodal data together with the novel challenges bring not just enriched understanding of tasks but also altogether new and fascinating opportunities for countless applications in the real world, such as autonomous driving, intelligent customer service provision and VR/AR etc.

The motivation behind the research of multimodal generative models primarily stems from their immense potential in information expression, task execution, and human-computer interaction. Compared to single-modal data, multimodal data can offer diverse perspectives on information, allowing models to gain a more comprehensive understanding of complex problems [4]. For instance, the combination of text and images enables machines to generate more precise descriptions of visual scenes, while the fusion of audio and video can facilitate a more natural interaction experience for virtual assistants. Additionally, multimodal generative models can perform information completion across different modalities, enhancing system robustness when handling incomplete or ambiguous inputs.

But they are not enough for effcient fusion in multimodal generation and have still some open issues to exploit the merits of multimodal data fully. The first one is the challenge of finding appropriate encoding methods for effective cross-modal alignment, due to nature and feature distribution differences across techniques. Second, multi-source information is fused to enhance multimodal generative models with powerful reasoning mechanisms that enable them to abstract essential features and generate plausible results. Finally, it is also important for the models to be trained in a way that they are scalable and can generalize well — particularly when targeting real-world data where we would like them to perform stably on multimodal inputs.

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

2 OVERVIEW OF MULTIMODAL GENERATIVE MODELS

2.1 Basic Concepts

Multimodal generative models perform a variety of complex tasks, such as translation, description generation, and content creation, by combining different modalities like text, images, and audio. These models can handle multiple input and output types, enabling transformations between different forms of data. For instance, tasks such as generating images from text or producing text descriptions from videos have shown significant effectiveness across various fields [5]. Through multimodal generative models, machines can understand and generate relevant information across modalities, greatly enhancing the quality and capability of generated content.

The recent rise of multimodal models like CLIP, DALL·E, and Flamingo has further showcased the immense potential of multimodal learning. These models leverage not only single-modal data but also fuse heterogeneous information from multiple sources to generate high-quality outputs. For example, the CLIP model, through joint training of images and text, comprehends the deep semantic relationships between them, leading to outstanding performance in tasks such as image description generation and image search. DALL·E demonstrates the ability to generate high-resolution images from text, creating high-quality visual content that aligns with complex textual descriptions. Models like Flamingo achieve closer interaction and information exchange among modalities such as images, text, and video through joint learning of multimodal data.

We demonstrate the power of multimodal generative models to advance content generation and interaction with information based on reciprocal fortes which have historically existed only in separate modalities but now come together — results that also hint at myriad future applications. Those models, using multimodal data effectively fuse contextual information and improve the accuracy as well as efficiency of completing a task on complicated scenarios.

2.2 Technological Evolution

The development of multimodal generative models can be traced back to early encoder-decoder frameworks, initially used for tasks like translation. These frameworks encoded input modalities into latent representations and then used decoders to generate outputs in other modalities. While this method laid the groundwork for multimodal generation, its effectiveness was constrained by the capabilities of single-modal information processing. With advancements in Generative Adversarial Networks (GANs) and large pretrained models, multimodal generative models have seen significant improvements. These technological innovations have made the conversion from one modality to another more precise and efficient, driving the development of multimodal generation tasks.

The emergence of GANs marked a breakthrough in the field of multimodal generation. GANs learn realistic data distributions through the adversarial process between a generator and a discriminator, achieving unprecedented results in tasks like image generation and image-to-text conversion. The generative capabilities of GANs are particularly well-suited for image generation tasks such as image translation and style transfer, showcasing strong potential in cross-modal generation. Variants like Conditional GANs and CycleGANs further extended the capabilities of multimodal generation by enabling models to generate outputs related to specified modalities through conditional inputs.

On the other hand, the rise of large pretrained models, particularly Transformer architectures and their variants—such as BERT, GPT, and T5—has significantly improved the performance of multimodal generative models. These models, pretrained on vast amounts of data, capture rich contextual information, enhancing understanding across multimodal data like text and images. Transformer-based multimodal generative models, such as DALL·E and CLIP, exhibit exceptional capabilities in generating images from text and producing textual descriptions from images. Pretrained models, leveraging cross-modal attention mechanisms, can better learn the relationships between different modalities, thus improving the quality and accuracy of generation.

The application of cross-modal attention mechanisms in multimodal generative models represents another key advancement. Attention mechanisms allow models to flexibly focus on relevant information sources when processing data from different modalities, thereby improving the coordination and conversion efficiency between modalities. This not only enhances model performance in multimodal generation tasks but also equips the model with stronger capabilities to comprehend complex cross-modal relationships. For instance, visual-language models can capture local features in images and associate them with textual descriptions through attention mechanisms, resulting in outputs that are more contextually relevant.

2.3 Model Classification and Mainstream Architectures

Multimodal generative models can be classified into various types based on the differences in input and output modalities, each corresponding to different tasks and application scenarios. Common classifications include text-to-image generation, image-to-text generation, and text-to-video generation. By integrating different modalities, these models can handle complex cross-modal tasks, achieving more natural and efficient content generation.

For example, text-to-image generation models like DALL·E generate high-quality images based on textual descriptions. These models can understand the details and

semantics within the text and transform them into visual content, making them widely applicable in creative design and visual presentation. DALL·E has demonstrated the ability to generate realistic and detailed images from complex textual prompts, marking a breakthrough in text-to-image generation tasks.

Conversely, image-to-text generation models, such as image description generation systems, extract visual features from input images to generate corresponding textual descriptions. Such systems hold significant application value in scenarios like image annotation, automatic interpretation of image content, and assisting visually impaired users. For example, by utilizing attention mechanisms, models can identify key information in images and generate contextually appropriate descriptions, improving the accuracy and semantic relevance of image-to-text conversions.

Mainstream multimodal architectures, such as CLIP and Flamingo, leverage powerful attention mechanisms to efficiently match modalities like text, images, and videos. CLIP achieves precise matching results in multimodal tasks, such as text-to-image retrieval and image-to-text generation, through joint training of text and images. Its core idea is to use attention mechanisms to establish deep semantic associations between text and images, allowing CLIP to perform excellently across various modal tasks through a shared representation space.

Flamingo further expands the application scope of multimodal models by integrating text, images, and video, providing capabilities for cross-modal generation and understanding. Flamingo utilizes hierarchical attention mechanisms to dynamically capture key features between text and visual information, enabling the generation of high-quality outputs. This type of model demonstrates significant potential in tasks such as video generation, complex scene description, and multimodal question answering.

3 APPLICATIONS OF MULTIMODAL GENERATIVE MODELS IN NLP

3.1 Integration of Audio and Language

1. Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) models convert speech signals into text form and have a wide range of applications in modern technology. Whether in everyday voice assistants like Siri and Google Assistant or in real-time transcription and meeting recording systems, ASR technology plays a crucial role. In recent years, deep learning-based models such as DeepSpeech and Wav2Vec 2.0 have significantly improved the accuracy and robustness of speech recognition. These models utilize deep neural networks and self-supervised learning techniques to reduce dependence on labeled data, enabling them to process complex speech inputs across various languages and accents, thus providing users with a smoother and more accurate voice interaction experience[6].

2. Text-to-Speech Synthesis (TTS)

Text-to-Speech (TTS) systems convert text into natural and fluent speech output. This technology is widely used in navigation systems, e-book reading assistants, and especially provides essential support for visually impaired individuals. Advanced TTS models like Tacotron 2 and WaveNet generate high-quality speech output that closely resembles human voice, achieving natural language synthesis[7]. These models can capture features such as tone, pauses, and pitch variations, making the synthesized speech more natural and expanding its application scenarios. For instance, TTS technology in voice navigation systems provides users with clear voice directions, significantly enhancing driving safety and user experience.

3. Music Generation and Speaker Recognition

Multimodal systems are not limited to speech recognition and synthesis but also show tremendous potential in music generation and speaker recognition. By integrating audio features with textual context, multimodal models can create unique musical content, bringing new possibilities to music creation and the entertainment industry. Additionally, speaker recognition technology analyzes feature information in speech to accurately identify different speakers, finding extensive applications in security and identity verification. This technology is particularly important in telephone customer service systems and intelligent security systems, effectively improving system security and user privacy protection.

3.2 Text Generation

Multimodal machine translation enhances translation quality by utilizing visual information such as images or videos, particularly demonstrating significant advantages in scenarios that require more precise contextual understanding. Traditional machine translation models rely on pure text input, while multimodal machine translation introduces visual modalities (like images or videos) to leverage visual context and improve language comprehension, thereby enhancing translation accuracy. Models like Vision-and-Language Pretraining (VLP) effectively capture the relationships between multimodal data by combining language and visual features during pretraining. Visual information can provide additional semantic cues and help the model resolve ambiguities that may exist in text, especially in language structures or scene descriptions with strong visual dependencies. Multimodal machine translation has been applied in image caption translation and video content translation, providing new pathways for cross-language content understanding.

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

Multimodal generative models demonstrate strong capabilities in summary generation and natural language generation. By combining text, images, audio, and other modalities, these models can extract key information from various types of data and generate structured text summaries. This is particularly useful for handling complex, multimodal data (such as richly illustrated news reports). Multimodal summary generation models can extract core information from non-text modalities like images and videos, generating concise text summaries that help users understand information more quickly.

4 APPLICATIONS OF MULTIMODAL GENERATIVE MODELS IN COMPUTER VISION (CV)

4.1 Image Generation and Editing

4.1.1 Text-to-Image Generation

Text-to-image generation models such as DALL·E and Stable Diffusion are able to generate novel photo-realistic images from text prompts. These are models capable of extracting all the semantic information found within a text and generating images corresponding to this description. An important application of Text-to-image generation technology is in the field such as artistic creation, design processes and advertising etc. The art and design processes involved are enhanced due to the nature of these tools, enabling artists and designers to visually explore their ideas quickly producing assets that can aid in realising creative concepts rapidly inspiring new ones. Secondly, manufacturers can adopt tailor-made images for bespoke market needs from the advertising industry to boost promotional effectiveness. As the quality of generation improves, text-to-image is being gradually used as a mainstream technique in commercial applications and personal creation.

4.1.2 Image-to-Image Translation

Image-to-image translation technology can enhance and process images through techniques such as style transfer and super-resolution reconstruction. These multimodal generative models show great potential in fields like medical imaging processing, image restoration, and artistic style conversion. By applying style transfer, models can transform images into representations of different artistic styles, creating new possibilities for artistic creation. Super-resolution reconstruction technology can significantly improve the clarity and detail of images, making blurred or low-resolution images clearer and more usable. In the medical imaging field, high-quality images are crucial for improving diagnostic capabilities; utilizing image-to-image translation technology allows doctors to better identify and analyze patient imaging data, leading to more accurate diagnoses.

4.1.3 Image Description Generation

Give a description of an ImageImage Description generation technology automatically created the required text descriptions for images. This is a great way to make things more accessible when using social media and for people who are visually impaired. By automatically attaching descriptions to photos, social platforms can vastly improve the accessible content available so that more users — including those who are partially sighted or blind — can access and understand image-based posts as intended. Conventional approaches employ Convolutional Neural Networks (CNN) to extract pixel-level important features and elements in the image, followed by a Recurrent Neural Network(RNN) Transformer structures for language generation that yields coherently the semantically meaningful text descriptions. These technological developments not only enhance human-computer interaction but also help to establish a digital world that can be more inclusive[8].

4.2 Video Generation and Editing

4.2.1 Video Generation

Video generation technology, based on spatiotemporal convolution and Transformer generative models, can generate and predict sequences of future frames. This technology has broad applications across various fields, including autonomous driving, surveillance, and entertainment. In autonomous driving, video generation can predict the future state of the environment around vehicles, assisting systems in real-time decision-making to ensure driving safety. In surveillance, video generation technology can generate continuous frames of scenes for anomaly detection and event prediction, enhancing the efficiency and accuracy of security monitoring[9]. In the entertainment sector, video generation can aid in content creation and animation production by generating high-quality video sequences, reducing production costs and accelerating creative cycles. As deep learning technology continues to advance, the quality and real-time capabilities of video generation are improving, driving the development of related applications.

4.2.2 Visual Content Enhancement and 3D Reconstruction

Visual content enhancement and 3D reconstruction technology mainly rely on super-resolution models and 3D object detection technology, playing a vital role in satellite imagery, medical analysis, and virtual/augmented reality (VR/AR). Super-resolution models can significantly improve image detail, making low-resolution images clearer, which is especially important in satellite image analysis, helping Geographic Information Systems (GIS) and environmental monitoring achieve more accurate data. In medical analysis, super-resolution technology can enhance the details of medical images, improving doctors' interpretative abilities

and thereby enhancing diagnostic outcomes.

Additionally, 3D object detection technology can accurately locate and identify objects in images, facilitating a more realistic interactive experience in VR/AR applications. In VR/AR environments, precise 3D reconstruction not only enhances user immersion but also improves spatial awareness, allowing users to interact and explore in virtual spaces more naturally[10]. The combination of these technologies is driving innovation across multiple fields, resulting in more efficient and intuitive visual experiences.

4.3 Physical Scene Simulation

Multimodal systems play a crucial role in simulating physical phenomena (such as fluids, flames, smoke, etc.), and these technologies have been widely applied in engineering design, educational research, and film special effects.

In engineering design, multimodal systems can provide high-precision simulations of physical phenomena, helping engineers optimize designs during product development. For instance, fluid dynamics simulations can be applied in aerospace, automotive engineering, and architectural design by thoroughly analyzing airflow and water flow behavior, ensuring that products achieve optimal performance and safety standards in actual use. This simulation technology not only increases design accuracy but also shortens development cycles and reduces costs.

In educational research, multimodal simulation technology provides students and researchers with intuitive learning and research tools. Through virtual laboratories, students can observe and experiment with complex physical phenomena, such as the formation and propagation of flames and smoke, deepening their understanding of related scientific principles. This interactive learning method not only stimulates student interest but also enhances learning effectiveness and promotes deeper scientific exploration.

In film special effects, the application of multimodal systems provides creators with powerful tools to generate high-quality visual effects. By simulating real physical phenomena like fire and smoke, film production teams can create more realistic scenes, enhancing audience immersion and viewing experience. These technologies enable modern movies, animations, and games to showcase unprecedented visual effects, providing audiences with more engaging storytelling experiences. Multimodal systems can optimize simulations of complex physical phenomena by leveraging advanced sensor data, thereby enhancing the accuracy and personalization of applications such as gait recognition[11].

5 Multimodal Generative Models Combining NLP and CV

5.1 Visual Question Answering (VQA)

Visual Question Answering (VQA) is an innovative multimodal generative model that combines visual data and textual data to answer questions related to the content of images or videos. The core of this model lies in understanding and analyzing the input visual information while generating corresponding answers based on the text questions posed by the user. VQA models typically utilize attention mechanisms to focus on areas of the image that are directly relevant to the question when processing complex visual information. This mechanism enables the model to extract the most valuable features, improving the accuracy of the answers.

The application scenarios of VQA technology are very diverse, covering various fields such as virtual assistants, educational tools, and social media. By integrating visual and linguistic elements, VQA models promote innovation in human-computer interaction, providing users with a more intuitive way to access information. Additionally, this technology lays the groundwork for understanding complex visual content and achieving smarter systems, demonstrating the immense potential of multimodal generative models in the fields of natural language processing and computer vision[12].

5.2 Multimodal Dialogue Systems

Multimodal dialogue systems can generate more context-aware responses by integrating images and text. The core of these systems lies in processing information from different modalities simultaneously, thereby providing more comprehensive and accurate answers. Unlike traditional single-modal dialogue systems, multimodal dialogue systems leverage the combination of visual information and textual input to enhance the depth and richness of conversations. By analyzing image content, these systems can capture subtle nuances related to user intent, allowing them to understand not only the surface meaning of language but also the context and emotional undertones. This capability enables the systems to make more appropriate responses in handling complex conversational scenarios and exhibit greater flexibility in dynamic interactions.

The application prospects of multimodal dialogue systems are very promising across multiple domains. For instance, in the online customer service sector, these systems can quickly identify and understand user inquiries by integrating images and text provided by users. This ability significantly enhances the accuracy and efficiency of customer service responses, allowing users to express problems visually rather than relying solely on textual descriptions. In the context of virtual assistants, such systems facilitate more natural interactions between humans and machines, enabling users to ask questions more flexibly and thereby enhancing overall user experience. Additionally, interactive learning systems utilizing multimodal dialogue technology can help students understand and analyze visual materials more effectively, promoting deep learning and knowledge absorption[13]. Through real-time feedback and guidance, students can receive more personalized support during their learning process, thereby improving learning

outcomes and engagement. With continuous technological advancements, multimodal dialogue systems will continue to drive the evolution of human-computer interaction, opening up new application areas and possibilities.

5.3 Emotion Recognition

Multimodal emotion recognition systems integrate multiple information sources, including voice, facial expressions, and body language, to provide more accurate and comprehensive emotional analysis. The design of these systems is based on a core principle: the complementary nature of different modalities that can reinforce one another. Voice information contains emotional cues such as tone, speed, and emphasis, while facial expressions directly reflect an individual's emotional state, and body language offers an additional dimension of underlying emotions. By comprehensively analyzing this diverse information, emotion recognition systems can accurately capture users' emotional fluctuations, avoiding misjudgments that may arise from relying on a single modality. This multi-layered information fusion not only enhances the precision of emotion recognition but also lays the foundation for subsequent emotional intervention and feedback mechanisms.

In terms of application, multimodal emotion recognition systems play a significant role in mental health assessment and user experience research. In the field of mental health, these systems can monitor and analyze individuals' emotional changes in real-time, providing valuable support to mental health professionals. By continuously tracking users' emotional states, the systems can help identify potential psychological issues and offer personalized intervention suggestions. In user experience research, multimodal emotion recognition systems can analyze users' emotional responses during product use, gathering authentic user feedback to provide crucial data support for product optimization. The application of this technology enables brands to gain deeper insights into consumer emotions, formulate more precise marketing strategies, and enhance user satisfaction and loyalty.

6 CASE STUDIES

6.1 Healthcare

In the healthcare sector, multimodal systems play a vital auxiliary role by integrating text and imaging data, particularly in medical image analysis and patient data generation. These systems can consolidate information from various sources to provide more comprehensive diagnostic support. By analyzing medical images (such as X-rays, CT scans, and MRIs) alongside relevant clinical text records, the systems can not only identify potential lesions in the images but also combine this information with patient histories and symptom descriptions, thus enhancing the accuracy and quality of diagnoses. This capability for multimodal fusion enables healthcare professionals to quickly access critical information and make more precise judgments, ultimately improving patient treatment outcomes.

Moreover, multimodal systems demonstrate unique advantages in protecting patient privacy. During the processing of medical data, these systems can combine imaging data with non-sensitive text information, reducing direct exposure of sensitive information and minimizing the risk of data breaches. Additionally, by employing advanced encryption techniques and data de-identification processes, these systems ensure effective protection of patient privacy. In terms of data generation, multimodal systems can utilize existing medical images and text data to create new simulated cases, which is significant for medical education, research, and model training.

6.2 Education and Training

The application of multimodal generative models in education is increasingly gaining attention, as these models can generate personalized teaching materials based on students' needs, offering a learning experience tailored to individual characteristics. By integrating text, images, videos, and other media, these models can automatically create content that meets the diverse learning styles and interests of different students. For example, for visual learners, the system can produce learning materials enriched with images and charts; for auditory learners, it can provide instructional videos with narrated audio. This personalized content generation not only enhances students' engagement but also improves their understanding and retention of information.

By integrating with virtual tutors or teaching assistants, multimodal generative models can also provide students with interactive learning experiences. These virtual assistants can respond to students' questions in real-time and adjust teaching strategies based on students' learning progress and feedback. Compared to traditional teaching methods, this interactive learning approach better addresses students' individual needs, offering immediate feedback and support. Furthermore, virtual assistants can analyze students' learning behaviors using multimodal data to identify challenges, thereby providing targeted tutoring. This education solution based on multimodal generative models not only enhances learning efficiency and enjoyment but also alleviates teachers' burdens, allowing them to focus more on individual differences and learning needs. As technology continues to advance, multimodal generative models will bring deeper transformations to the education sector, promoting diverse and personalized learning methods.

6.3 Entertainment and Creative Industries

Multimodal generative models are gradually expanding into the entertainment industry, especially in game generation—creating complex gaming environments—and film-making for special effects with remarkable possibilities of using them to enhance user experience. By combining

various types of data, such as images, audio and text; models are able to build more complex worlds (environments) which are experienced by users. For example, multimodal generative models in gaming can start scaled game environments and characters while also create new interactions on the fly. The technology not only means reductions in development time, but also enables the creation of individual experiences wherein an endlessly evolving virtual world can be explored and interacted with.

In the realm of VR content generation, multimodal models can merge elements from the real world with virtual environments to create immersive experiences for users. These systems can dynamically adjust the composition and interactive elements of the virtual environment based on users' real-time feedback and behaviors, enabling freer exploration and participation in immersive experiences. This flexibility and adaptability make the creation of VR content more personalized and interactive, enhancing users' sense of engagement and satisfaction.

In film special effects production, multimodal generative models also play a crucial role. By integrating visual effects with audio elements, these models can generate highly realistic scenes and actions, adding visual impact and emotional depth to films. With these technologies, film production teams can create complex special effects shots more quickly, reducing production costs while improving the quality of the final product. This cross-modal creative capability not only enriches the forms of entertainment content but also drives innovation within the industry. As technology continues to advance, multimodal generative models will increasingly play a pivotal role in the entertainment field, providing users with deeper immersive experiences.

7 CHALLENGES AND FUTURE DIRECTIONS

7.1 TECHNICAL CHALLENGES

Despite the broad application prospects of multimodal generative models in various fields, several key technical challenges remain. Firstly, the scarcity of multimodal datasets limits the training and performance enhancement of these models. Compared to unimodal data, the complexity of collecting and annotating multimodal data makes it difficult to obtain high-quality datasets, which in turn affects the generalization ability of the models. Secondly, the effectiveness of modality fusion is also a pressing issue that needs to be addressed. The differences in data characteristics between different modalities pose a core challenge in multimodal generative model research—how to effectively fuse these modalities to extract meaningful information. Additionally, as model scales continue to expand, computational cost has become an important consideration. Large-scale models have high hardware resource requirements during training and inference, which limits their application in resource-constrained environments.

7.2 SECURITY ISSUES

Generative models can be maliciously exploited, such as for creating deep fake content, which may involve fake news, videos, or audio. Once such false content is disseminated, it can seriously impact social opinion, political stability, and personal privacy. Therefore, establishing effective detection and review mechanisms to ensure the authenticity and safety of generated content is an urgent issue that needs to be resolved. Specifically, technical developers need to research and implement content detection algorithms to differentiate between real and fake content, thereby reducing potential harm to society. Furthermore, multimodal generative models must prioritize data security when processing user data. When collecting and utilizing multimodal data (such as user voice, images, and behavioral data), ensuring data encryption, secure storage, and compliant use is fundamental to building a trustworthy multimodal system. Data breaches or misuse not only jeopardize user privacy but can also lead to legal liabilities. Therefore, developing secure data processing workflows and technologies to prevent unauthorized access or tampering of data is an essential measure to ensure user safety[14]. Additionally, the security and robustness of the system must complement each other during the development and deployment of multimodal generative models. Developers need to consider potential attack methods, such as adversarial attacks, to ensure that models maintain stable performance against various threats. This includes validating input data to prevent malicious inputs from affecting the generated content.

Lastly, establishing corresponding security standards and regulations is crucial. Various stakeholders, including technology developers, policymakers, and the public, must work together to promote responsible technology application. This involves not only regulating the technology itself but also enhancing public awareness of the potential security risks associated with its use, fostering vigilance and coping capabilities. Only on the foundation of technological security and responsible usage can multimodal generative models truly unleash their potential to bring about positive societal changes.

7.3 FUTURE RESEARCH DIRECTIONS

Looking ahead, research on multimodal generative models will develop in multiple directions to address current technical challenges and meet the growing application demands. Firstly, developing lightweight models is key to enhancing the adoption rate. With the proliferation of the Internet of Things (IoT) and mobile devices, users increasingly demand real-time and efficient AI solutions. By optimizing model structures and algorithms, such as model pruning, quantization, and knowledge distillation, it is possible to significantly reduce computational resource consumption. This will not only enable real-time inference on edge and mobile devices but also expand the model's

applications in low-power environments, benefiting more users.

Secondly, improving model interpretability is an important future research direction. As AI technologies become widely adopted, users' understanding and trust in the model's decision-making process are increasingly important. A transparent decision-making process can help users understand the reasons behind the model's outputs, thus enhancing their trust and reliance on the model. To achieve this, researchers can explore interpretability techniques, such as attention mechanism visualization, model proxy methods, and backpropagation interpretability, to reveal the internal mechanisms of models when processing multimodal inputs. This interpretability will not only enhance user trust but also provide insights for analysis and improvement when errors occur.

Moreover, enhancing the performance of multimodal generative models in human-computer interaction will lay the foundation for a more natural and efficient user experience. Traditional human-computer interaction often relies on unimodal information input, while multimodal systems can offer a richer and more intuitive interactive experience by integrating text, voice, and visual information. Researchers can focus on intelligent interaction methods that enable systems to understand users' intentions and emotions, thereby providing more personalized feedback. For example, based on users' historical behavior, contextual information, and emotional states, models can generate responses that are more aligned with users' needs, further improving user satisfaction. The use of multimodal generative models plays a vital role in this method; medical applications can benefit significantly from integration with BCI (brain-computer interface) technology[15]. For example, BCIs can sense the neural activity of patients while multimodal systems analyze visual and text data to offer real-time feedback during medical examinations. Clinicians can achieve this by linking vital signs with visual diagnostics through a new integration point[16]. When investigating alterations in the uterocervical angle or cervical length during pregnancy and their predictive value for spontaneous preterm birth according to various studies, combining these markers allows for a comprehensive view of patients' health. More broadly, in complex case analyses involving multimodal data-mining models, this approach is likely to lead to improved information retrieval and, ultimately, more effective interventions and care plans[17]. This innovative strategy marks a new era in the diagnosis and treatment of child health issues, demonstrating how we can leverage technological tools to identify patient-centered phenotypic states, resulting in more effective therapeutic outcomes.

By promoting research in these areas, the full application potential of multimodal generative technology in education, healthcare, and entertainment can be realized. For instance, large language models, like those explored in Efficient Fine-Tuning, can enhance personalized learning materials to improve student outcomes in education[18]. In healthcare, accurate diagnostic tools can leverage multimodal data to improve patient treatment quality. Meanwhile, in the entertainment sector, these technologies can create more engaging and immersive content for users.

8 CONCLUSION

This paper has explored the extensive applications of multimodal generative models in natural language processing (NLP) and computer vision (CV), emphasizing their transformative impact across various industries. By integrating multiple modalities such as text, images, and audio, these models not only enhance the expression and understanding of information but also advance the depth of human-computer interaction. In the healthcare field, multimodal generative models can assist doctors in analyzing imaging data and clinical information more accurately, improving diagnostic efficiency; in education, they provide personalized learning experiences, fostering improved learning outcomes; and in entertainment, these models create more immersive and interactive user experiences, reshaping the production of games and film content.

As technology continues to advance, multimodal generative models will further drive the development of AI-driven systems, delivering more intuitive, efficient, and accessible solutions. Future research will focus on addressing current technical challenges such as dataset scarcity, the effectiveness of modality fusion, and computational costs. Simultaneously, ethical and societal issues will become important focal points to ensure the safe and responsible use of technology. Through continuous optimization and innovation, multimodal generative models will achieve broader applications across multiple sectors such as healthcare, education, and entertainment, promoting sustainable societal development.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

LI, Liang

Shandong Youth University of Political Science, China.

REFERENCES

[1] Huang, X., Wu, Y., Zhang, D., Hu, J., & Long, Y. (2024). Improving Academic Skills Assessment with NLP and Ensemble Learning. arXiv preprint arXiv:2409.19013.

[2] Ma, B., Ma, B., Gao, M., Wang, Z., Ban, X., Huang, H., & Wu, W. (2021). Deep learning - based automatic inpainting for material microscopic images. Journal of Microscopy, 281(3), 177-189.

[3] Liu, W., Cheng, S., Zeng, D., & Qu, H. (2023). Enhancing document-level event argument extraction with contextual clues and role relevance. arXiv preprint arXiv:2310.05991.

[4] Wang, D. (Ed.). (2016). Information Science and Electronic Engineering: Proceedings of the 3rd International Conference of Electronic Engineering and Information Science (ICEEIS 2016), January 4-5, 2016, Harbin, China. CRC Press.

[5] Liu, W., Zhou, L., Zeng, D., Xiao, Y., Cheng, S., Zhang, C., ... & Chen, W. (2024). Beyond Single-Event Extraction: Towards Efficient Document-Level Multi-Event Argument Extraction. arXiv preprint arXiv:2405.01884.

[6] Lu, J. (2024). Optimizing E-Commerce with Multi-Objective Recommendations Using Ensemble Learning.

[7] Yu, P., Cui, V. Y., & Guan, J. (2021, March). Text classification by using natural language processing. In Journal of Physics: Conference Series (Vol. 1802, No. 4, p. 042010). IOP Publishing.

[8] Jiang, L., Yang, X., Yu, C., Wu, Z., & Wang, Y. (2024, July). Advanced AI framework for enhanced detection and assessment of abdominal trauma: Integrating 3D segmentation with 2D CNN and RNN models. In 2024 3rd International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC) (pp. 337-340). IEEE.

[9] Wang, Y., Ban, X., Wang, H., Li, X., Wang, Z., Wu, D., ... & Liu, S. (2019). Particle filter vehicles tracking by fusing multiple features. IEEE Access, 7, 133694-133706.

[10] Wang, C., Kang, D., Sun, H. Y., Qian, S. H., Wang, Z. X., Bao, L., & Zhang, S. H. (2024). MeGA: Hybrid Mesh-Gaussian Head Avatar for High-Fidelity Rendering and Head Editing. arXiv preprint arXiv:2404.19026.

[11] Bačić, B., Feng, C., & Li, W. (2024). JY61 IMU SENSOR EXTERNAL VALIDITY: A FRAMEWORK FOR ADVANCED PEDOMETER ALGORITHM PERSONALISATION. ISBS Proceedings Archive, 42(1), 60.

[12] Qu, M. (2024). High Precision Measurement Technology of Geometric Parameters Based on Binocular Stereo Vision Application and Development Prospect of The System in Metrology and Detection. Journal of Computer Technology and Applied Mathematics, 1(3), 23-29.

[13] Zhang, Y., Wang, F., Huang, X., Li, X., Liu, S., & Zhang, H. (2024). Optimization and Application of Cloud-based Deep Learning Architecture for Multi-Source Data Prediction. arXiv preprint arXiv:2410.12642.

[14] Cao, Y., Weng, Y., Li, M., & Yang, X. The Application of Big Data and AI in Risk Control Models: Safeguarding User Security.

[15] Liu T, Wu Y, Ye A, Cao L, Cao Y. Two-stage sparse multi-objective evolutionary algorithm for channel selection optimization in BCIs. Frontiers in Human Neuroscience. 2024 May 22;18:1400077.

[16] Zhang, M., Liu, Y., Zhang, B., Li, S., & Yu, H. (2024). Unilateral complete ureteral duplication with ectopic ureteral opening inserting into urethra in a female patient without incontinence: a case description and review of the literature. Quantitative Imaging in Medicine and Surgery, 14(8), 6166172-6166172.

[17] Zhang, M., Li, S., Tian, C., Li, M., Zhang, B., & Yu, H. (2024). Changes of uterocervical angle and cervical length in early and mid-pregnancy and their value in predicting spontaneous preterm birth. Frontiers in Physiology, 15, 1304513.

[18] Leong, H. Y., Gao, Y. F., Shuai, J., Zhang, Y., & Pamuksuz, U. (2024). Efficient Fine-Tuning of Large Language Models for Automated Medical Documentation. arXiv preprint arXiv:2409.09324.