Journal of Computer Science and Cybernetics, V.42, N.1 (2026), 1–24
DOI no. 10.15625/1813-9663/23021

MULTIMEDIA MULTIMODAL ARTIFICIAL INTELLIGENCE (MMAI): FOUNDATIONS, CHALLENGES, AND FUTURE DIRECTIONS

LE HOANG SON[1*], ONI DAMILOLA[2]

[1] *VNU Information Technology Institute, Vietnam National University, 144 Xuan Thuy Street, Cau Giay Ward, Ha Noi, Viet Nam*
[2] *International School, Vietnam National University, HT1 Building, VNU Campus at Hoa Lac, Ha Noi, Viet Nam*

Abstract. Multimedia Multimodal Artificial Intelligence (MMAI) represents a transformational paradigm that enables machines to process and synthesize many modalities, e.g., text, image, audio, and video, to understand and generate complex multimedia content. This review provides an intensive exploration of multimedia multimodal artificial intelligence, which focuses on its basic models, major challenges, and future directions. Drawing insight from recent research trends and literature, this paper presents a comprehensive analysis of multimodal AI, fusion techniques, self-supervised learning strategies, and real-world applications such as healthcare, education, entertainment, and human-computer interactions. It also examines the theoretical foundation of MMAI, including multimodal representation, alignment, and fusion techniques, which are very important to integrate heterogeneous data sources while maintaining coherence and relevance. The review also mentions the role of self-supervised learning in reducing dependence on labeled datasets by taking advantage of the underlying structure of multimodal data. Additionally, this review highlights the ability of generic AI to create multimedia content, stretching the limits of what AI can do in creative and practical domains. Despite this progress, many challenges persist, including technical limitations like high computational costs, data inequality or heterogeneity, and model interpretability, as well as ethical concerns relating to privacy and bias. Finally, future research directions will be mapped out, including the development of scalable and efficient training methods, the integration of symbolic reasoning with deep learning, and the promotion of interdisciplinary collaboration. By synthesizing knowledge from leading studies and industry innovations, this review will be a blueprint for people, which aims to exploit the full potential of AI-driven multimedia technologies in an increasingly interconnected world.

Keywords. AI 4.0, content generation, fusion techniques, human-computer interaction, multimedia, multimodal AI, self-supervised learning.

1. INTRODUCTION

Multimedia Multimodal Artificial Intelligence (MMAI) represents a transformative pattern where AI systems process and synthesize multiple information or data, such as text, images, audio, video, and sensor readings, to understand and produce complex multimedia content [1]. The idea of multimodal AI began to take shape in late 2010. One of the initial

*Corresponding author.
E-mail addresses: sonlh@vnu.edu.vn (L.H. Son), igbagbooluwadamilola@gmail.com (O. Damilola).

attempts was the use of Deep Boltzmann Machines jointly for modeling images and text in 2014. While demonstrating the viability of integrating various data types, it faced boundaries in scalability and performance due to the complexity of Deep Boltzmann Machines [2]. In 2015, the show-and-tell model marked an important milestone using a long-term short-term memory (LSTM) network for generating text descriptions and utilizing a convolutional neural network (CNN) for image feature extraction [3]. This model greatly improved the quality of image captioning, highlighting the capacity of sequence models such as LSTMs in multimodal tasks. Visual Question Answering (VQA) models emerged around 2016, combining a CNN for image understanding with a recurrent neural network (RNN) for question processing. These models enabled complex arguments on visual and text data, although their performance depended heavily on the embedding quality from both modalities. The introduction of a transformer-based model brought about a new era in multimodal AI. As the first effort, VilBERT (2019) extended BERT to handle visual and text input using individual streams for each model, setting a new benchmark at various benchmarks [4]. However, these models were computationally intensive. In 2020, VisualGPT combined GPT-2 with a visual encoder to generate coherent text based on the visual input, further demonstrating the ability of generative transformers in multimodal learning [5]. Recent trends in multimodal AI include the development of integrated models such as OpenAI GPT-4 Vision and Gemini of Google, which can handle text, images, and other data types within the same architecture. Nevertheless, the development of multimodal AI in terms of multimedia applications, emphasizing its basic models, key challenges, and way forward is still under the progress. Drawing insights from recent literature and research trends, this review provides an analysis of real-world applications in domains such as multimodal generative models, fusion techniques, self-inspection education, ethical considerations, healthcare, education, and entertainment. Ultimately, this review acts as a roadmap for researchers who aim to use the full potential of AI-driven multimedia technologies in an interconnected world. The diagram below provides an overview of the main contents of this article.

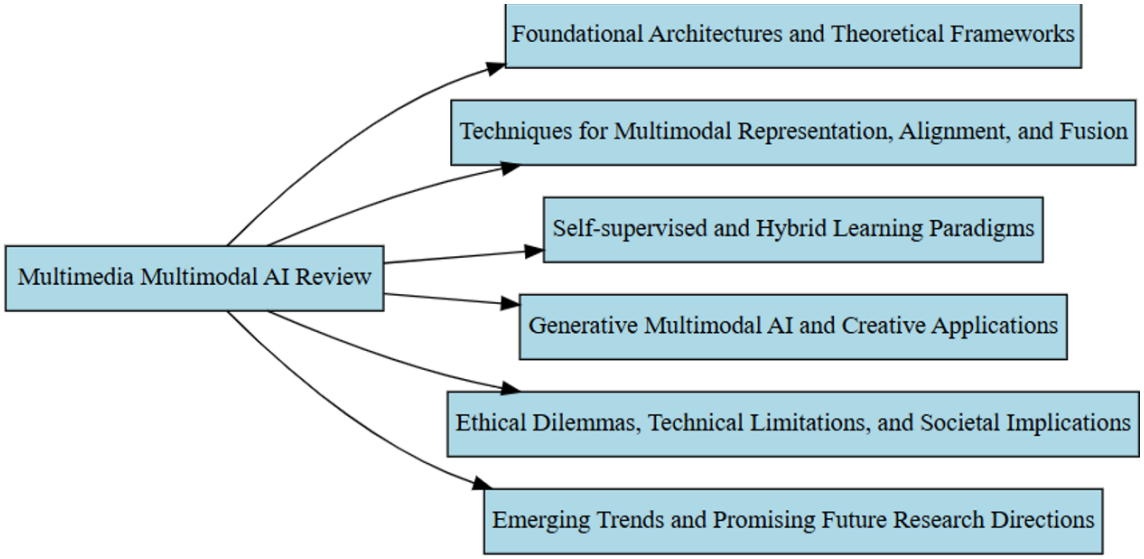

Figure 1: overview of the paper structure

2. AI FOR MULTIMEDIA INTEGRATION

2.1. Advancement of multimedia and AI integration

Integration of Multimedia and Artificial Intelligence has evolved significantly in the last few years, marked by important milestones shaping the landscape of modern technology. Initially, the focus was on developing algorithms with the ability to process individual media types like text, pictures, audio, and video separately. The early attempts to combine these modalities were rudimentary, often limited by computational obstacles and an extensive dataset deficiency. However, as computing power increased and data availability expanded, researchers began to detect the ability to integrate many forms of media to create a stronger and more versatile system.

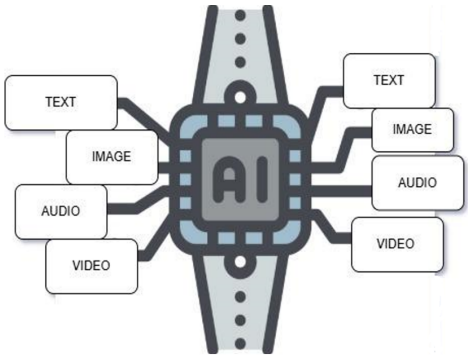

Figure 2: Multimodal AI

During 1980s and early 1990s, the concept of multimedia computing emerged, emphasizing the need for systems that could handle different data types simultaneously. This period saw the introduction of the basic multimedia framework, which allows for synchronization of audio and video, paving the way for more complex interactions [6]. As the Internet gained popularity in the mid-1990s, the amount of multimedia content exploded, inspiring the development of search engines and recommendation systems to suit multimedia data. These systems rely more on metadata and simple feature extraction techniques, laying the groundwork for more advanced AI applications[7].

The turn of the millennium marked an important shift in the integration of multimedia and AI, especially with the development of machine learning techniques. Researchers began to apply statistical learning strategies to multimedia analysis; this made way for more advanced classification and retrieval tasks. Notably, using Hidden Markov Models and Support Vector Machines (SVM) became common in areas like image recognition and speech processing[8]. This also brought the rise of content-based retrieval structures, which aimed to understand the semantics of multimedia data rather than simply indexing it based on metadata.

As the 2000s progressed, the integration of multimedia and AI took a huge step forward with introduction of deep learning techniques. The convolutional neural network (CNN) revolutionized image processing and reached unprecedented accuracy in works like object detection and image classification [9]. In addition, the recurrent neural network (RNN) enhanced the capabilities of natural language processing, allowing a refined understanding of text and speech. These progresses facilitated the creation of multimodal systems, which

process and synthesize information with different modalities [10]. This further leads to huge innovations in areas such as automated captioning, sentiment analysis, and even augmented reality applications.

An important milestone in this development came up with the development of a multimodal learning framework, seeking to unify the processing of diverse data types within a single model. Techniques like Multimodal Compact Bilinear Pooling (MCBP) and Cross-modal Attention Mechanisms (CMAM) emerged, enabling the system to learn joint representations, which captured the relationship between various modalities. These frameworks laid the foundation for more advanced applications, such as visual question answering (VQA) and cross-model retrieval, where the performance of the system depends a lot on understanding the interplay between text and images [11].

Recent years have witnessed an increase in interest in generative AI models, especially those that can create new materials across modalities. Tools like GANs (Generative Adversarial Networks) and VAES (Variational Autoencoders) have enabled the synthesis of realistic images, videos, and audio, which opens new avenues for creative expression and content creation. These approaches go beyond traditional limits of what AI can do but can also raise important questions about the ethical implications of generating synthetic media and authenticity.

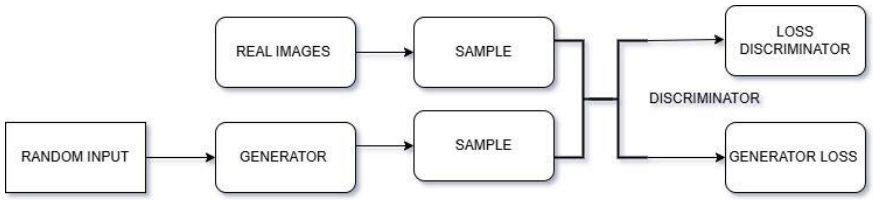

Figure 3: GAN structure

The historical context of multimedia and AI integration reveals a dynamic difference between technological advancements and developing user needs. From the early days of separate media processing to the current scenario of integrated intelligent systems, each stage has contributed to the rich tapestry of possibilities that define today's multimedia experiences. As we delve deeper in the potential of multimedia multimodal AI, it is necessary to reflect on this history to inform future innovations to ensure they align with social values with ethical ideas.

2.2. Theoretical foundations of multimedia multimodal AI

The MMAI is a set of fundamental principles at the core of artificial intelligence, which controls how diverse modalities like text, images, audio, video, and sensor data can be processed and aligned, and also fused to create intelligent systems that are capable of understanding and generating complex multimedia materials [12]. These principles form the theoretical backbones of multimedia multimodal artificial intelligence, which guide the development of models that can effectively integrate heterogeneous or odd data sources while maintaining consistency, coherency, and contextual relevance. The most important concepts in this domain are multimodal representation, alignment, and fusion. Each will play a distinguished role in enabling the AI system to process and synthesize multimodal information.

2.2.1. Multimodal representation: Encoding heterogeneous data into unified spaces

Multimodal representation is the procedure of converting raw data from various modalities into a shared embedding space where meaningful comparisons and interactions may occur [13]. Since each modality, either raw text, visual, or audio possesses its unique characteristics and structure, direct comparison or integration is challenging. Therefore, the goal of multimodal representation is to capture the high-level semantic characteristics that abstract away from modality-specific intricacies, preserving the essential meaning conveyed by each input type [14].

There are two primary approaches for constructing multimodal representations: joint representations and coordinated representations. In joint representations, features from all modalities are added to a single vector space through operations such as concatenation, element-wise summation, or bilinear pooling. This allows for the overall processing of multimodal data within an integrated structure, making it useful for tasks like visual question answering (VQA) and multimodal sentiment analysis. On the other hand, coordinated representations maintain separate embeddings for each modality but apply similarity constraints to ensure that inputs related to sequences from different types of modalities are close to their respective space[15]. Recent progress in deep learning has introduced more sophisticated techniques for multimodal representation, including attention-based models and transformer architecture. The attention mechanisms allow the model to provide dynamic importance weights to different parts of an input sequence; this enables selective focus on the most relevant features across modalities [16]. Transformer-based models, especially multimodal variants such as vision transformers (VITS) and cross-modal transformers, have further enhanced the ability to capture long-range dependence between modalities, facilitating more accurate and contextually aware representations [17].

2.2.2. Multimodal alignment

While multimodal representation focuses on encoding diverse inputs in a shared space, multimodal alignment is relative to identifying clear relationships between elements from various modalities. This is important for the tasks that require precise coordination between modalities, such as matching the words spoken with the corresponding video segment, adding captions with relevant image areas, or synchronizing gestures with speech. Effective alignment ensures that the multimodal model can accurately map components from one modality to its counterparts in the second, enhancing understanding and reasoning abilities [5]. Alignment is broadly classified into explicit and implicit alignment. Implicit alignment occurs when models automatically infer correspondence during training without requiring predefined notes [18].

For example, attention-based models can implicitly align modalities, dynamically serving vital parts of a modality by processing another. This is commonly observed in image caption systems, where the model learns to associate specific image regions with corresponding words in the generated caption. Explicit alignment, on the other hand, involves predefined mapping or notes that guide the alignment process. This approach is usually use in supervised settings, where labeled data provides direct supervision to align modalities, such as in data sets containing paired images and descriptions with region-word annotations [19].

Several methodologies have been proposed to improve alignment accuracy, including

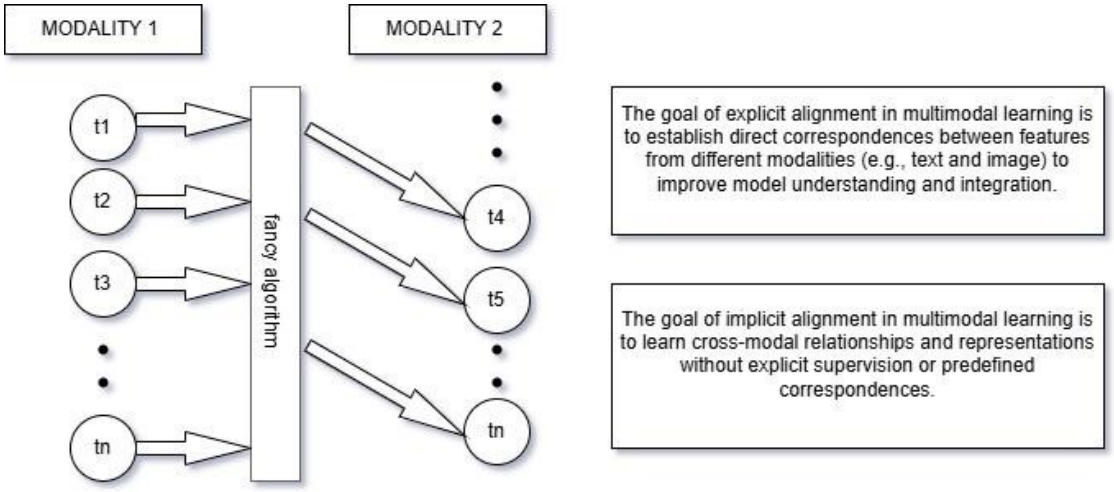

Figure 4: Multimodal alignment

cross-modal mediation networks, graph-based models, and contrastive contractive learning techniques. Lacey et al. [20] used Cross-modal Attention Networks to use multi-head attention mechanisms to establish fine-grained associations between different modalities, allowing models to focus on the most prominent features when creating predictions. Yang [21] also used Graph-based models to represent multimodal data as a structured graph, where nodes correspond to entities (e.g., an object in image or words in a sentence) and the edges denote the relationship between them. Contrastive learning techniques, inspired by the paradigms of self-supervised learning, train models to differentiate between positive, negative pairs of multimodal samples and reinforce correct alignments while discouraging mismatches [22].

2.2.3. Multimodal fusion: Integrating information for task-specific outcomes

Once alignment has been obtained after multimodal representation is established, the final phase in the multimodal AI pipeline is fusion, including integrating information from multiple modalities to produce the task-specific output. Fusion can occur at different levels of abstraction, which depend on the nature of application and the desired result. Broadly, there are three main fusion strategies: early fusion, late fusion, and hybrid fusion [23]. Early fusion combines raw features from various modalities before feeding into a downstream model. This approach assumes that integrating the modalities in an early stage allows it to learn a joint representation that captures interaction between modalities from the outset. However, early fusion can be computationally expensive and can introduce noise if irrelevant or redundant features are included [24]. Late fusion, on the contrary, processes each modality freely by using different models and then adds its output at the latter stage, usually through voting mechanisms, ensemble methods, or learned combination functions [25]. Hybrid fusion seeks to create a balance between these extremes by incorporating both early and late fusion components, benefiting the models from both low-level feature interactions and high-level decision fusion [26].

Advanced fusion techniques have been developed to increase the effectiveness of multimodal integration. Bilinear fusion methods, such as multimodal compact bilinear pooling, calculate pair interactions between various features from various modalities, capturing high-order relationships that can be omitted by simple fusion approaches[27]. Transformer-based

fusion models leverage self-attention mechanisms to dynamically change the contributions of each modality on contextual cues, and it enables adaptive fusion that adapts to the specific requirements of each given task [28].

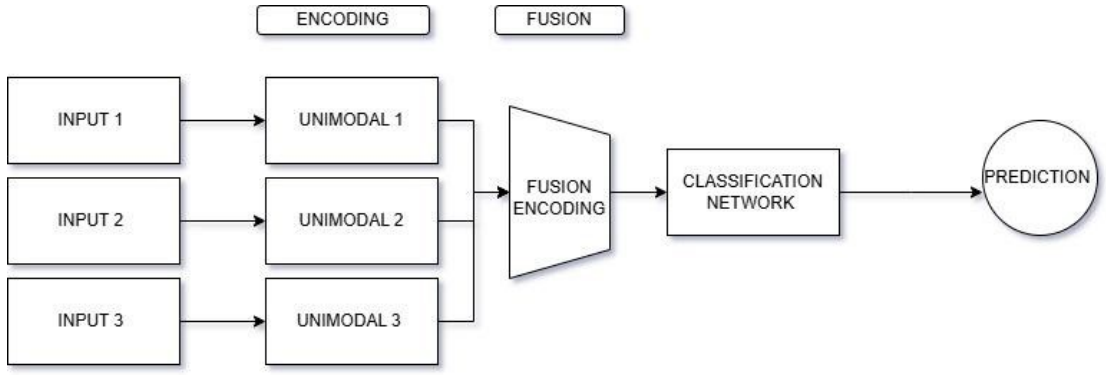

Figure 5: Architecture for Intermediate/Late Multimodal Fusion

Additionally, neural architecture search techniques are automatically employed to discover optimal fusion strategies to suit specific applications, which improves performance [29]. By combining these fundamental principles, representation, alignment, and fusion, multimedia multimodal artificial intelligence can get a deep understanding of multimodal data, allowing them to carry out complex tasks with higher accuracy and higher efficiency [30].

2.3. Key architectures and models in multimedia multimodal AI

The rapid growth of MMAI has been driven by the growth of sophisticated deep learning architectures that are able to process and synthesize multiple modalities together [31]. The most impressive models are GANs, VAES, transformers-based framework, and hybrid they integrate symbolic arguments with deep learning. Each of these architectures plays a clear and recognizable role in advancing multimodal AI, which enables successes in content generation, cross-modal logic, and intelligent decision-making [32].

2.3.1. Generative adversarial networks (GANs): Realistic content generation across modalities

Generative Adversarial Network (GAN), introduced by Goodfellow et al. (2014) [33], has revolutionized the generative AI field by enabling highly realistic images, videos, and even audio sequences. At its core, GANs include two competitive neural networks: a generator, which learns to make synthetic data, and a discriminator, which evaluates the authenticity of the output generated [33]. Through adversarial training, the generator improves its ability to produce content that looks like real-world data, while discriminator refines its ability to distinguish between real and synthetic samples. In terms of MMAI, GANs have been enhanced to handle multimodal inputs, allowing the generation of content that integrates several modalities [34]. For example, text-to-image GANs, such as StackGAN and AttnGAN, generate photorealistic images from text descriptions by leveraging attention mechanisms to align linguistic elements with the visual features. Similarly, Audio-Visual GANs always synthesize synchronized video sequences with audio input to enable applications such as music-powered video generation and lip-sync animation [35]. Recent progress,

such as StyleGAN-V, has improved the controllability and quality of the output generated, making GANs an important cornerstone of multimodal content creation [36].

2.3.2. Variational autoencoders (VAEs): Probabilistic latent representations for multimodal reconstruction

Variational autoencoders (VAEs), introduced by Kingma and Welling (2013) [37], a chance-based method or statistical approach to generative modeling by learning latent representations of data distributions. Unlike deterministic autoencoders, VAEs impose a predistribution in the latent space, which allows constant projection between stochastic sampling and data points [37]. This makes VAEs well-suited for the tasks that require structured generations, such as denoising, inpainting, and style transfer.

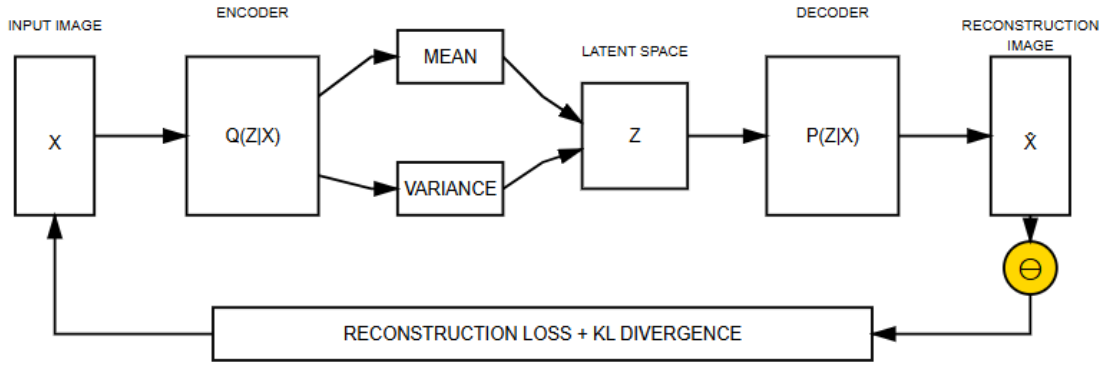

Figure 6: Variational autoencoder architecture

In MMAI, VAEs have been adapted to handle multimodal data by expanding their latent space to accommodate multiple input types. Multimodal VAEs encode information in a shared latent representation from different modalities, which enables cross-modal reconstruction and translation [38]. For example, a text-image VAE may rebuild an image from a text description or generate a descriptive caption from the input image.

2.3.3. Transformers: Attention-based models for joint multimodal processing

Transformers architecture was introduced in 2017 by Vaswani et al. [39], became dominant architecture in natural language processing. These caused wide adoption in multimodal AI, facilitating joint processing of diverse modalities [39]. Unlike traditional recurrent or convolutional networks, Transformers works on sequences of the tokens and weigh the importance of different elements through attention heads, thereby enabling flexible and scalable multimodal reasoning [40].

In MMAI, Vision Transformer (VITs) has been employed to process images by shuffling the sequence of patches, which allows for spontaneous integration with text data in multimodal settings [41]. Models such as CLIP (contrastive language–image pretraining) take advantage of a large-scale dataset for learning joint embedding of images and text, which enables zero-shot transfer in tasks like image classification and cross-modal retrieval [41]. Additionally, cross-modal transformers have been developed to align and fuse information

from various modalities, supporting applications like visual question answering and video captioning. The scalability and adaptation of transformers make them a powerful tool for producing end-to-end multimodal AI systems [42].

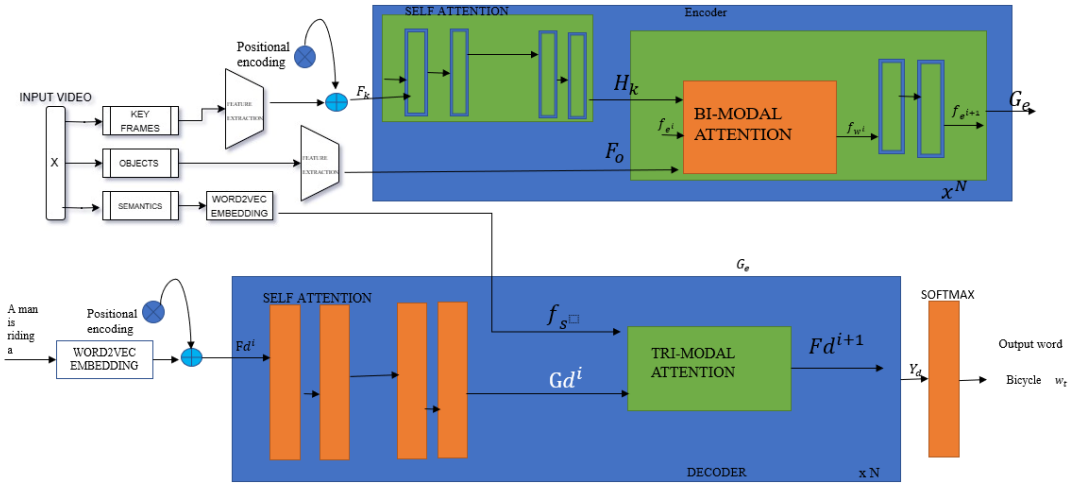

Figure 7: Transformers architecture

While deep learning models excel in pattern recognition and data-driven learning, they often misinterpret logical reasoning abilities. To address these boundaries, researchers used hybrid models that integrate symbolic reasoning with deep learning. These models combine the strengths of the rules-based systems and neural networks, which enable AI to perform tasks that require both perceptual understanding and high-level reasoning [43]. A prominent example is the use of Knowledge graphs in combination with deep learning models. Knowledge graphs provide a structured representation of factual knowledge, allowing AI to reason about relations between system entities [44]. When combined with a multimodal model, a knowledge graph system can make more accurate inferences by grounding the predictions of deep learning in semantic knowledge. Additionally, neural-symbolic AI approaches include logic-based reasoning in neural architectures, which enables tasks such as explainable AI and commonsense reasoning. For example, in healthcare diagnostics, hybrid models can integrate medical ontologies with deep learning-based symptom analysis to improve clinical accuracy and transparency [45]. Another rising trend is the development of modular neural architectures, where various components are specialized in specific subtasks and collaborate through interpretable interfaces. Models like Differentiable Neural Computer and Neural Turing Machine include model memory modules that allow the AI system to store and retrieve information dynamically, mimicking human logic [46]. The diversity of deep learning architecture available today has greatly expanded the capabilities of multimedia multimodal artificial intelligence, enabling improvement of intelligent systems that process, understand, and generate multimodal content with unprecedented sophistication. GANs and VAEs have pushed the boundaries of generative AI to transformers that facilitate joint multimodal reasoning and contribute uniquely to the advancement of each architectural field. In addition, hybrid models that integrate symbolic arguments with deep learning provide a promising path towards more explanatory and logically consistent AI systems.

3. SELF-SUPERVISED LEARNING

3.1. Self-supervised learning in multimedia multimodal AI

Self-supervised learning is a powerful paradigm in MMAI, addressing the challenges of data scarcity and reduced dependence on annotated datasets [47]. Unlike traditional supervised learning, which depends on widely labeled data, self-supervised learning uses built-in relationships between different types of data, e.g., text, images and audio, to teach models without needing labeled examples i.e., it enables the creation of meaningful training signals without manual labeling [48].

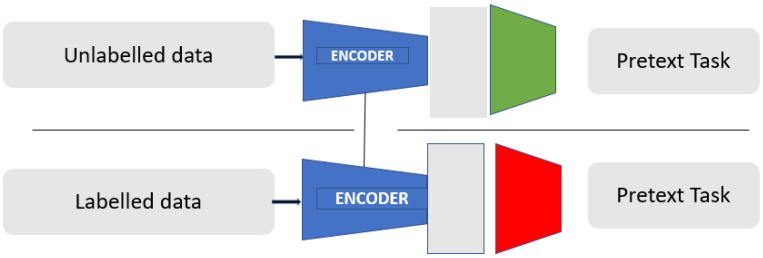

Figure 8: Simple schematic representation of self supervised learning paradigm

This allows it to learn by exploiting relationships between several modalities and eventually improves performance in a wide range of tasks like cross-modal retrieval, visual question answering, and generative content synthesis. Several self-supervised learning strategies in the multimedia multimodal artificial intelligence have gained prominence, including masked modeling, contrastive learning, and cross-modal prediction, each of which contributes uniquely to the advancement of multimodal understanding and logic.

3.1.1. Masked modeling: Reconstructing missing inputs for representation learning

Masked modeling is a widely adopted self-supervised learning technology that trains the model to recreate the missing or corrupted parts of input data based on relevant information from other modalities [48]. Inspired by the success of BERT (Bidirectional Encoder Representations from Transformers) in natural language processing (NLP), masked modeling has been extended to multimodal settings, where models learn to predict masked segments in a modality using information from others [49]. For example, masked autoencoder and BEiT (Bidirectional Encoder Representation from Transformers with Image Tokenization) apply masking to visual input, re-forming occluded image regions using visible ones. In multimodal contexts, this theory is applied across modalities so that models can fill the gaps in a modality using data from another.

A remarkable implementation of masked modeling in multimedia multimodal artificial intelligence is FLAVA (Fusion of Vision and Language pre-trained models with Alignment), which jointly masks text and image tokens and trains the model to reconstruct them using a cross-modal context [50, 51]. Similarly, VideoMAE applied masked modeling to video data, rebuilding missing frames based on surrounding temporal and spatial information [52]. One of the advantages of masked modeling lies in its ability to learn hierarchical features

that capture both local and global dependence, which makes it effective for tasks requiring fine-grained reasoning

3.1.2. Contrastive learning: Enhancing discriminative capabilities through similarity maximization

Contrastive learning is another fundamental self-supervised learning strategy that amplifies the discriminative power of multimodal models by encouraging the same representation for positive sample pairs while creating distance with dissimilar ones [53]. This approach depends on making positive-negative pairs, where positive pairs have semantically related multimodal samples (e.g., an image and its related captions), and negative pairs are unrelated samples.

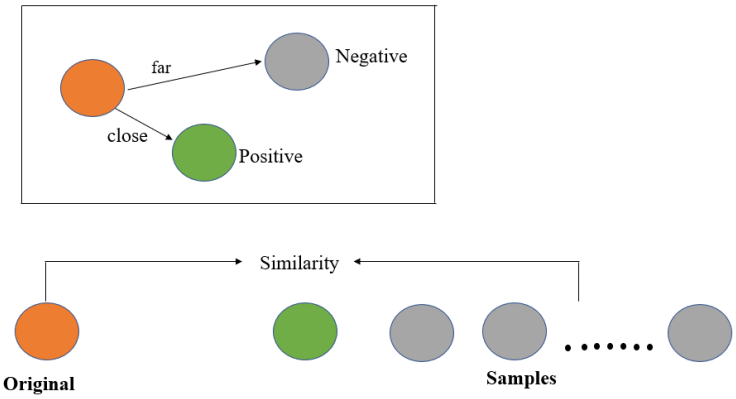

Figure 9: Visualization of how Contrastive learning works

Prominent examples of contrastive learning in multimedia multimodal artificial intelligence include Momentum Contrast (MoCo) and Simple Framework for Contrastive Learning of Visual Representations (simCLR), which have been adapted for multimodal settings [54]. In the case of CLIP (Contrastive Language–Image Pretraining), the model is trained on a massive dataset of image-caption pairs, which learns to align visual and text representations in shared embedding space. The success of CLIP shows how contrastive learning can bridge the gap between different modalities, which can enable zero-shot transfer in downstream tasks like image classification and cross-modal recovery. Additionally, InfoNCE (Noise Contrastive Estimation) loss, a variant of contrastive learning often used to train models on large-scale multimodal datasets, refines their ability to differentiate between relevant and irrelevant information [55]. The strength of contrastive learning depends on its capability to produce highly discriminative embeddings, especially for tasks that require accurate alignment between modalities. However, its effectiveness is contingent on the quality of positive-negative pairs it constructs; it requires careful design of augmentation strategies and sampling techniques to avoid biases and ensure generalization [56].

3.1.3. Cross-modal prediction: Leveraging inter-modality dependencies for supervised signal generation

Cross-modal prediction is self-supervised learning technology that takes advantage of the inherent relationships between different modalities to generate supervisory signals without manual annotation [57]. This approach trains the model to predict a modality from each

other and effectively uses a modality as a proxy for supervision, e.g, we can train a model to generate text details of an image or to synthesize the image from a given text, allowing it to learn bidirectional mappings between modalities.

One of the initial examples of cross-model prediction in multimedia multimodal artificial intelligence is DeViSE (deep visual-semantic embedding model), which trains a model to predict word embeddings from image features, aligning visual and text representation in a shared space [58]. Recently, the Aligning Vision and Language has refined this approach by scaling up training on large web-crawled datasets, gaining a cutting edge level of cross model retrieval and image classifications [59]. Additionally, video-text models such as VideoBERT and CBOW (continuous bag-of-words) expand the cross-modal prediction for temporal data and learn to generate text summaries from video sequences or vice versa [60].

Cross-model prediction provides many advantages, including labeled or noisy data and the potential for zero-shot generalization. However, it also presents challenges related to modality mismatch and ambiguity, as some text details can suit many visual interpretations and vice versa [61]. To address these challenges,a careful design of loss functions and architectural choices is required to ensure accurate consistency between predicted and actual modalities.

3.2. Advantages of self-supervised learning in MMAI

Adopting self-supervised learning in multimedia multimodal artificial intelligence provides several benefits, especially to reduce dependence on labeled datasets and improve model generalization.

Firstly, self-supervised learning reduces the bottleneck of manual annotation, which is expensive and time-consuming in multimodal settings where several modalities must be aligned. Taking advantage of unlabeled data enables it learning model to learn from an enormous amount of freely available multimedia content, which expands its applicability to domains with limited annotated resources [62].

Secondly, it enhances the strength of the model by exposing it to various data variations and encouraging it to capture underlying structural patterns. This leads to better generalization across unseen tasks and domains, making self-supervised learning-based models more adaptable to real-world scenarios. Also, self-supervised learning serves as the development of foundation models, which serve as pre-trained backbones for a wide range of downstream tasks. Models like CLIP, ALIGN, and FLAVA exemplify how self-supervised learning can be used for the creation of universal multimodal representations that effectively move to applications like image captioning, visual question answering, and cross-modal search [63].

Finally, self-supervised learning supports the integration of emerging modalities beyond text, image, and audio, such as haptic feedback, physiological signals, and environmental sensors. By enabling the model to learn from diverse input sources without clear supervision, self-supervised learning paves the way for inclusive and multimodal-aware AI systems that are compatible with the real world [64].

4. APPLICATIONS OF MMAI

The versatility and adaptability of multimedia multimodal artificial intelligence enable its broad applications in various domains, which are transforming industries such as creative content production, healthcare, human-computer interactions, and education. By integrating many forms, such as text, image, audio, and video, multimedia multimodal artificial

intelligence systems can provide more comfortable, personal, and immersive experiences. Below, we discuss some of the most effective applications of multimedia multimodal artificial intelligence in these fields, showing that they are shaping the workflows, enhancing decision-making, and redefining user experience.

4.1. Creative content generation: Redefining art, music, and storytelling

One of the most captivating edges of MMAI is its ability to generate creative content that rival human-made content in originality and creativity. Advances in generative models like GANs, Variational Autoencoders, and Transformers enabled artificial intelligence to create engrossing visual art, compose songs, and craft narratives that engage audiences in innovative ways [65].

In the purview of visual art, models such as DALL·E, Stable Diffusion, and Midjourney have demonstrated notable proficiency in generating high-resolution images from text prompts. Artists and designers now use these to detect new aesthetic styles, rapidly prototype ideas, and even collaborate with AI to create artworks [66]. Beyond static images, multimedia multimodal artificial intelligence also ventures into animated and cinematic content generation, where models such as Make-a-Video and Meta's Make-A-Scene translate textual descriptions into short video clips, opening new possibilities for filmmakers and digital storytellers.

Beyond the auditory and visual creativity, multimedia multimodal artificial intelligence is also changing literature and storytelling. Large language models like GPT-4 and LLaMA can generate coherent and relevant rich stories, while multimodal story generators such as StoryGAN and NarrativeFlow craft interactive stories from the combination of textual and visual elements [67]. These tools are being used in gaming, advertising, and educational content creation, where personalized storytelling enhances user engagement and immersion.

4.2. Healthcare: Enhancing diagnostics, treatment, and patient engagement

The healthcare sector has greatly benefited from the integration of MMAI, especially in areas such as diagnostic imaging, clinical decision support, patient monitoring, and personal treatment schemes. Multimedia multimodal artificial intelligence systems gives detailed, accurate diagnostic framework and care distribution, combining medical imaging, electronic health records, lab results, and patient-reported symptoms [68].

One of the most prominent applications of MMAI in healthcare is medical imaging analysis, where AI models assist radiologists in detecting abnormalities such as tumors, fractures, and lesions. Models like IBM Watson Health and Google's DeepMind Health use multimodal fusion techniques to integrate MRI, CT scan, and histopathological data with the history of the patient to improve clinical accuracy and reduce false positives [69]. Additionally, pathology analysis benefits from multimedia multimodal artificial intelligence, where deep learning models trained on a digitized biopsy slide can detect cancer cells with high precision and thus augment the work of pathologists.

Beyond imaging, MMAI plays an important role in clinical decision support systems, where AI analyzes the multimodal data stream, including significant signs, genetic markers, and behavioral indicators to recommend treatment plans. For example, platforms such as Tempus and Pathway Genomics use AI to personalize cancer remedies based on genomic profiles, Delivering a personalized treatment strategies that align with each patient's unique biological profile [70].

Patient engagement and remote monitoring have also been revolutionized by MMAI-powered conversational agents and virtual health assistants [71]. Chatbots such as Babylon Health and ADA Health integrate NLP with symptom-checking algorithms to provide initial diagnoses and triage recommendations. Meanwhile, wearable devices equipped with AI-powered analytics track physiological data like sleep patterns, heart rate and activity levels to detect conditions like diabetes, cardiovascular diseases and mental disorders.

4.3. Human-computer interaction: Enabling immersive and intuitive interfaces

Human-computer interactions have been deeply influenced by MMAI, causing the development of intelligent virtual assistants, augmented reality (AR)/virtual reality (VR) systems, and emotion-aware interfaces that enhance user experience and access. Intelligent virtual assistants like Amazon Alexa, Google Assistant, and Apple Siri uses multimodal fusion techniques to process voice commands, visual inputs, and contextual cues and to provide more natural and responsible interactions. In immersive technologies, multimedia multimodal artificial intelligence plays a vital role in multimodal artificial intelligence AR and VR applications, where real-time tracking, gesture recognition, and visual recognition are essential. Companies like Meta (formerly Facebook) and Microsoft are taking advantage of AI to increase mixed reality experiences, allowing users to interact more comfortably with digital content [72]. For example, AI-powered avatars in the VR environment can simulate realistic facial expressions and gestures, which can improve social appearance in virtual meetings and collaborative workspaces.

Additionally, emotion-aware computing is receiving traction as a means to create adaptive and empathetic interfaces. MMAI models can detect emotional stages in real time, trained on facial expressions, physiological signals, and voices, allowing applications to adjust their behavior accordingly [73].

4.4. Education: Personalized learning and intelligent tutoring systems

There are some education learning platforms that are personalized and are powered by MMAI, such as Khan Academy and Coursera, that use multimodal data, including students' performance metrics, engagement levels, and behavioral cues, for tailored content distribution. AI-operated tutors such as Carnegie Learning's MATHia and Squirrel AI analyze learners' progress in real-time and provide customized clarifications, practice, and reactions to optimize comprehension and retention. The interactive learning system further enhanced the engagement by incorporating natural language processing (NLP) and computer vision to facilitate conversational learning. To generate dynamic exercises based on the user's response, platforms like Duolingo and Quizlet leverage AI to ensure that learners receive targeted instructions that are compatible with their proficiency level. Additionally, AI-operated translation tools assist students with disabilities or language barriers, making educational material more accessible [74].

In addition, MMAI is being used to develop smart classrooms, where AI-powered analytics monitor the engagement, detect learning difficulties, and provide real-time intervention. Smart whiteboards, AI-assisted grading systems, and automated proctoring tools are streamlining administrative functions, allowing teachers to focus more on teaching and mentorship [75]. As the AI models continue to develop, their effect will only deepen, unlocking new possibilities for innovation and changes across industries.

5. CHALLENGES AND FUTURE DIRECTIONS

5.1. Challenges in multimedia multimodal AI

Despite the remarkable progress, it faces several important challenges that affect its widespread adoption and effectiveness. These challenges stem from the complexity of diverse data methods, the computational demands of multimodal models, and the complexity of integrating ethical dilemmas around data privacy and AI-generated content. Understanding and addressing these issues is important to ensure that multimedia multimodal artificial intelligence systems are technically strong, socially responsible, and trustworthy.

5.1.1. Technical challenges: High computational costs and data heterogeneity

One of the primary technical challenges in MMAI is high computational cost, which is associated with training and deploying multimodal models. Unlike unimodal AI systems, which process data from a single source, the multimodal model must handle multiple input streams, such as text, images, audio, and video, each requiring separate preprocessing and feature extraction techniques. This complexity increases the number of parameters in deep learning models, leading to prolonged training time, high energy consumption, and greater hardware requirements [76]. For example, large-scale multimodal models like Contrastive Language–Image Pretraining and Aligning Vision and Language require extensive computational resources to process the billions of image-caption pairs, which makes them impractical for many real-world applications with limited infrastructure.

In addition, data heterogeneity is an important challenge in learning multimodality. Different modalities exhibit different formats, resolutions, and temporal dynamics, making them difficult to align and integrate into a unified representation [77]. For example, while text data is discrete and sequential, visual and audio data are continuous and high-dimensional. To bridge these differences, sophisticated feature engineering and normalization techniques are required, which can introduce noise or lose important information. Additionally, the presence of missing or incomplete methods in the real-world dataset complicates training, as it will learn to handle partial input without compromising performance. To address these challenges, more efficient architecture is required, such as lightweight multimodal transformers and modular neural networks that can dynamically change to various input configurations [78].

5.1.2. Ethical and societal concerns: Privacy, bias, and misinformation

Beyond the technical limitations, MMAI raises ethical and social concerns, especially about privacy, prejudice, and misinformation. For example, a facial identification system that combines biometric data with behavioral patterns can reveal personal attributes such as identity, emotional status, or even health status without clear consent. Similarly, multimodal monitoring systems that analyze both visual and audio inputs pose a risk for individual freedom, especially when deployed in public places without regulatory oversight. Ensuring data anonymity, implementing strict access controls, and following ethical AI guidelines are important steps towards reducing these risks [79].

Bias in the AI model is another important concern in MMAI. Training data often reflects existing societal prejudices, which leads to improper consequences in applications such as hiring, law enforcement, and healthcare diagnostics [80]. For example, trained multimodal

models on biased datasets can display racial or gender inequalities in facial recognition accuracy or medical diagnosis. A multifaceted approach is required to address bias, including diverse and representative dataset curation, fairness-aware training techniques, and post-processing adjustments to reduce discrimination effects.

The rise of generative AI in MMAI has further increased the concerns about misinformation and deepfakes. Advanced models such as DALL·E, Stable Diffusion, and StyleGAN-V can produce extremely realistic images, videos, and audio that are extremely similar to the authentic contents [81]. While these abilities have creative and commercial applications, they also cause serious risk in terms of fake news, political manipulation, and identity fraud. The ease with which AI-generated media can be manipulated can underline the need for strong detection mechanisms, watermarking techniques, and a legal framework to regulate the use of synthetic contents.

5.1.3. Interpretability and Trustworthiness in AI decision

Many multimodal models, particularly those based on deep learning, serve as "black boxes" that conceals the logic behind decisions making [82]. This lack of transparency can destroy the user's confidence, especially in high sectors like healthcare, finance, and criminal justice. e.g., if a system recommends a medical treatment based on multimodal data, physicians must be able to find out the reason behind the recommendation to assess its validity. For example, the attention mechanism in the transformer has the ability to highlight which parts of an input (e.g., specific image areas or words in a sentence) influenced a model prediction.

Features such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) provide insights into how individual input components contribute to the final output. Additionally, hybrid models that combine deep learning with symbolic reasoning provide more transparent decision-making routes, which allows for human-in-loop verification. Benchmarking initiatives such as the Multimodal Evaluation Workshop and the Vision-And-Language Reasoning Challenge offer standardized tests to evaluate model performances in real-world scenarios. These efforts promote development of accountable AI systems that are ethical and technical standards [83].

5.2. Conclusion

As MMAI continues to develop, many promising research directions are emerging, aimed at addressing current boundaries and expanding the capabilities of the multimodal system. These include model architecture progression, scalable and highly efficient training methods, responsible AI development, and interdisciplinary collaboration amongst stakeholders. By pursuing these routes, researchers will enhance the performance, interpretability, and moral alignment of MMAI, which will ensure its permanent growth in diverse domains.

5.2.1. Advancements in model architectures

Future development in MMAI will be powered by innovations in model architecture that improve multimodal understanding, logic, and generations. A major direction is the integration of more modalities, including haptic feedback, olfactory signals, and physiological data, to create rich and more immersive AI experiences. Current models mainly focus on text, images, audio, and video, but expanding these capabilities to include additional sensory

inputs can be more comprehensive and context-aware AI systems. For example, in healthcare, integrating bio signals such as electrocardiograms or brainwave readings with visual and text data may increase clinical accuracy and personalized treatment plans.

Although deep learning excels at pattern recognition, it often lacks the interpretability required for high-stakes applications such as legal reasoning, scientific discovery, and autonomous decisions [84]. Including symbolic logic reasoning in neural architecture will ensure the AI model performs logical deductions, causal inference, and commonsense thinking, making them more transparent and reliable. Approach such as neuro-symbolic AI, which merge the neural network with knowledge graphs and rules-based systems, are already showing promise in visual question answering and medical diagnosis.

Additionally, dynamic and adaptive architecture is expected to gain more traction, allowing models to adjust their processing pipelines based on input characteristics. Modular neural networks, where various components are specialized in specific functions and collaborate, provide a promising solution. Models like Differentiable Neural Computer (DNC) and Neural Turing Machine (NTM) show how the system can save and recollect information dynamically, which mimics human-like reasoning. Extending these concepts to multimodal settings will lead to flexible and context-aware models that can handle complex, real-world scenarios [85].

5.2.2. Scalable and efficient training methods

As multimodal models grow in complexity, demand for scalable and efficient training methods becomes rapidly important. Current cutting edge models like CLIP, ALIGN, and FLAVA require massive computational resources to train on a large-scale dataset, making them inaccessible to many organizations [86]. To address this, researchers are exploring methods like lightweight architecture, transfer learning, and meta-learning to improve model efficiency without waiving performance.

Lightweight models, such as TinyML and MobileNet variants, aim to reduce computational overhead by optimizing neural network structures for edge computing. These models can run on mobile devices and embedded systems, enabling real-time multimodal processing in a low-resource-constrained environment. Additionally, knowledge distillation, where small models learn from large, pre-trained teachers, has shown promise in compressing the complex multimodal model while maintaining its predictive capabilities [87]. Transfer learning and meta-learning approaches are also receiving traction as a way to improve model adaptability. The transfer learning allows the learning model to take advantage of pre-trained representatives from a domain and fine-tune them for new tasks, reducing the need for extensive label data. Meta-learning, or "learning to learn", allows the model to adapts quickly to new tasks with little examples, which makes them ideal for few-shot and zero-shot learning scenarios. These techniques are particularly valuable in sectors like healthcare and education, where data deficiency is a frequent challenge.

Efforts are also underway to improve distributed training and optimization; researchers examined parallel computing techniques, federated learning, and hardware accelerators like GPUs and TPUs to speed up training. Optimizations such as gradient compression, sparse training, and quantization-aware learning are being explored to reduce memory and computational costs.

5.2.3. Responsible AI development

With increase influence of AI, its paramount to ensure responsible AI development. Data privacy remains another important concern, especially in applications involving individual or sensitive information. Differential privacy techniques, add noise to training to protect personal identity, are being investigated to safeguard user data. In addition, regulatory compliance and policy development are important for controlling and governing the use of multimedia multimodal artificial intelligence. Governments and organizations are working to establish guidelines for AI governance, ensuring that models follow legal and ethical standards. Initiatives like the European Union's AI Act and OECD AI Principles provide a foundation for responsible AI development, promoting transparency, accountability, and human inspection in the AI system.

5.2.4. Interdisciplinary collaboration

Finally, the future of MMAI depends on interdisciplinary collaboration, making experts from different fields come together to deal with complex challenges. AI researchers, cognitive scientists, policymakers, ethicists, and domain experts will work together to develop models that match human needs.

The future model of multimedia multimodal AI lies in constant continued innovation in model architecture, training efficiency, ethical development, and interdisciplinary cooperation. By addressing these, researchers can unlock the new possibilities for AI-driven multimedia technologies, ensuring their responsible and impactful integration into society.

REFERENCES

[1] T. Zubatiuk and O. Isayev, "Development of multimodal machine learning potentials: Toward a physics-aware artificial intelligence," *Accounts of Chemical Research*, vol. 54, no. 7, pp. 1575–1585, Apr. 2021.

[2] L. Bravo, C. Rodriguez, P. Hidalgo, and C. Angulo, "A systematic review on artificial intelligence-based multimodal dialogue systems capable of emotion recognition," *Multimodal Technologies and Interaction*, vol. 9, no. 3, p. 3, Mar. 2025.

[3] A. Mosavi, S. Ardabili, and A. R. Várkonyi-Kóczy, "List of deep learning models," in *Engineering for Sustainable Future*. Cham: Springer International Publishing, 2020, pp. 202–214.

[4] L. Li, G. Chen, H. Shi, J. Xiao, and L. Chen, "A survey on multimodal benchmarks: In the era of large ai models," *arXiv*, Sep. 2024.

[5] Y. Ektefaie, G. Dasoulas, A. Noori, M. Farhat, and M. Zitnik, "Multimodal learning with graphs," *Nature Machine Intelligence*, vol. 5, no. 4, pp. 340–350, Apr. 2023.

[6] "MPEG-21: The 21st century multimedia framework - IEEE journals & magazine - IEEE xplore," Online, accessed: May 22, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/1184339/

[7] D. Tjondronegoro and A. Spink, "Web search engine multimedia functionality," *Information Processing & Management*, vol. 44, no. 1, pp. 340–357, Jan. 2008.

[8] N. Zheng and J. Xue, *Statistical Learning and Pattern Analysis for Image and Video Processing*. Springer Science & Business Media, 2009.

[9] A. Mathew, P. Amudha, and S. Sivakumari, "Deep learning techniques: An overview," in *Advanced Machine Learning Technologies and Applications*. Singapore: Springer, 2021, pp. 599–608.

[10] S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Information Fusion*, vol. 36, pp. 10–25, Jul. 2017.

[11] "Visual question answering: a state-of-the-art review - artificial intelligence review," Online. [Online]. Available: https://link.springer.com/article/10.1007/s10462-020-09832-7

[12] "Integrated multimodal artificial intelligence framework for healthcare applications - npj digital medicine," Online. [Online]. Available: https://www.nature.com/articles/s41746-022-00689-4

[13] A. Gaonkar, Y. Chukkapalli, P. J. Raman, S. Srikanth, and S. Gurugopinath, "A comprehensive survey on multimodal data representation and information fusion algorithms," in *2021 International Conference on Intelligent Technologies (CONIT)*, Jun. 2021, pp. 1–8.

[14] S.-F. Zhang, J.-H. Zhai, B.-J. Xie, Y. Zhan, and X. Wang, "Multimodal representation learning: Advances, trends and challenges," in *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, Jul. 2019, pp. 1–6.

[15] Y. Chen, X. Ge, S. Yang, L. Hu, J. Li, and J. Zhang, "A survey on multimodal knowledge graphs: Construction, completion and applications," *Mathematics*, vol. 11, no. 8, p. 8, 2023.

[16] "Multibench: Multiscale benchmarks for multimodal representation learning - pmc," Online. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11106632/

[17] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, Oct. 2023.

[18] H. Hu *et al.*, "Learning implicit feature alignment function for semantic segmentation," in *Computer Vision – ECCV 2022*. Cham: Springer Nature Switzerland, 2022, pp. 487–505.

[19] T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao, and C. Sun, "An end-to-end textspotter with explicit alignment and attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5020–5029. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/He_An_End-to-End_TextSpotter_CVPR_2018_paper.html

[20] S. Lacey, A. Peters, and K. Sathian, "Cross-modal object recognition is viewpoint-independent," *PLOS ONE*, vol. 2, no. 9, p. e890, Sep. 2007.

[21] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo, "Deep multimodal representation learning from temporal data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5447–5455. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Yang_Deep_Multimodal_Representation_CVPR_2017_paper.html

[22] "A survey on contrastive self-supervised learning," Online. [Online]. Available: https://www.mdpi.com/2227-7080/9/1/2

[23] P. Rashinkar and V. S. Krushnasamy, "An overview of data fusion techniques," in *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Feb. 2017, pp. 694–697.

[24] G. Barnum, S. Talukder, and Y. Yue, "On the benefits of early fusion in multimodal representation learning," *arXiv*, Nov. 2020.

[25] "Effective techniques for multimodal data fusion: A comparative analysis," Online. [Online]. Available: https://www.mdpi.com/1424-8220/23/5/2381

[26] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, Sep. 2017.

[27] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv*, Sep. 2016.

[28] G. Zheng *et al.*, "A transformer-based multi-features fusion model for prediction of conversion in mild cognitive impairment," *Methods*, vol. 204, pp. 241–248, Aug. 2022.

[29] J.-M. Perez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "Mfas: Multimodal fusion architecture search," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 6959–6968.

[30] B. Jaltotage, J. Lu, and G. Dwivedi, "Use of artificial intelligence including multimodal systems to improve the management of cardiovascular disease," *Canadian Journal of Cardiology*, vol. 40, no. 10, pp. 1804–1812, Oct. 2024.

[31] V. Radu *et al.*, "Multimodal deep learning for activity and context recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 4, pp. 157:1–157:27, Jan. 2018.

[32] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

[33] I. Goodfellow *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020.

[34] E. P. Blasch, U. Majumder, T. Rovito, and A. K. Raz, "Artificial intelligence in use by multimodal fusion," in *2019 22th International Conference on Information Fusion (FUSION)*, Jul. 2019, pp. 1–8.

[35] V.-N. Nguyen, M. Sadeghi, E. Ricci, and X. Alameda-Pineda, "Deep variational generative models for audio-visual speech separation," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*, Oct. 2021, pp. 1–6.

[36] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A review on generative adversarial networks: Algorithms, theory, and applications," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3313–3332, Apr. 2023.

[37] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv*, Dec. 2013.

[38] "Joint variational autoencoders for multimodal imputation and embedding - nature machine intelligence," Online. [Online]. Available: https://www.nature.com/articles/s42256-023-00663-z

[39] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[40] "Transfg: A transformer architecture for fine-grained recognition - proceedings of the aaai conference on artificial intelligence," Online. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/19967

[41] "A review of generalized zero-shot learning methods - IEEE Journals & Magazine - IEEE Xplore," Online. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9832795

[42] S. Frank, E. Bugliarello, and D. Elliott, "Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers," *arXiv*, Sep. 2021.

[43] "AI reasoning in deep learning era: From symbolic AI to neural–symbolic AI," Online. [Online]. Available: https://www.mdpi.com/2227-7390/13/11/1707

[44] J. Xu, K. Chen, X. Qiu, and X. Huang, "Knowledge graph representation with jointly structural and textual encoding," *arXiv*, Dec. 2016.

[45] "Jmir medical informatics - identifying clinical terms in medical text using ontology-guided machine learning," Online. [Online]. Available: https://medinform.jmir.org/2019/2/e12596/

[46] J. Pérez, J. Marinković, and P. Barceló, "On the turing completeness of modern neural network architectures," *arXiv*, Jan. 2019.

[47] "Artificial intelligence in multimodal learning analytics: A systematic literature review - sciencedirect," Online. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666920X25000669

[48] "Self-supervised learning methods and applications in medical imaging analysis: a survey [peerj]," Online. [Online]. Available: https://peerj.com/articles/cs-1045/

[49] D. Mizrahi *et al.*, "4M: Massively multimodal masked modeling," *Advances in Neural Information Processing Systems*, vol. 36, pp. 58 363–58 408, Dec. 2023.

[50] D. Hong, C. Li, B. Zhang, N. Yokoya, J. A. Benediktsson, and J. Chanussot, "Multimodal artificial intelligence foundation models: Unleashing the power of remote sensing big data in earth observation," *Innovation in Geosciences*, vol. 2, no. 1, p. 100055, 2024.

[51] S. El-Sappagh, J. M. Alonso, S. M. R. Islam, A. M. Sultan, and K. S. Kwak, "A multilayer multimodal detection and prediction model based on explainable artificial intelligence for alzheimer's disease," *Scientific Reports*, vol. 11, no. 1, p. 2660, Jan. 2021.

[52] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 078–10 093, Dec. 2022.

[53] "Leveraging multimodal learning analytics to differentiate student learning strategies — proceedings of the fifth international conference on learning analytics and knowledge," Online. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/2723576.2723624

[54] P. Hager, M. J. Menten, and D. Rueckert, "Best of both worlds: Multimodal contrastive learning with tabular and imaging data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 924–23 935. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023/html/Hager_Best_of_Both_Worlds_Multimodal_Contrastive_Learning_With_Tabular_and_CVPR_2023_paper.html

[55] "[2008.05659] what should not be contrastive in contrastive learning," Online. [Online]. Available: https://arxiv.org/abs/2008.05659

[56] "Survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging," Online. [Online]. Available: https://www.mdpi.com/1099-4300/24/4/551

[57] "Artificial intelligence for multimodal data integration in oncology: Cancer cell," Online. [Online]. Available: https://www.cell.com/cancer-cell/fulltext/S1535-6108(22)00441-X

[58] A. Singh *et al.*, "FLAVA: A foundational language and vision alignment model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15 638–15 650. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Singh_FLAVA_A_Foundational_Language_and_Vision_Alignment_Model_CVPR_2022_paper

[59] "A comprehensive review of the video-to-text problem - artificial intelligence review," Online. [Online]. Available: https://link.springer.com/article/10.1007/s10462-021-10104-1

[60] T. H. Kappen, W. A. van Klei, L. van Wolfswinkel, C. J. Kalkman, Y. Vergouwe, and K. G. M. Moons, "Evaluating the impact of prediction models: lessons learned, challenges, and recommendations," *Diagnostic and Prognostic Research*, vol. 2, no. 1, p. 11, Jun. 2018.

[61] J. Gui *et al.*, "A survey on self-supervised learning: Algorithms, applications, and future trends," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9052–9071, Dec. 2024.

[62] B. Wallace and B. Hariharan, "Extending and analyzing self-supervised learning across domains," in *Computer Vision – ECCV 2020*. Cham: Springer International Publishing, 2020, pp. 717–734.

[63] "Self-supervised learning: Generative or contrastive - IEEE Journals & Magazine - IEEE Xplore," Online. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9462394

[64] "Project muse - artificial intelligence and visual art," Online, accessed: May 29, 2025. [Online]. Available: https://muse.jhu.edu/pub/6/article/599831/summary

[65] M. Mohammadi, E. Tajik, R. Martinez-Maldonado, S. Sadiq, W. Tomaszewski, and H. Khosravi, "Artificial intelligence in multimodal learning analytics: A systematic literature review," *Computers and Education: Artificial Intelligence*, p. 100426, May 2025.

[66] L. R. Soenksen *et al.*, "Integrated multimodal artificial intelligence framework for healthcare applications," *npj Digital Medicine*, vol. 5, no. 1, pp. 1–10, Sep. 2022.

[67] H.-P. Chan, R. K. Samala, L. M. Hadjiiski, and C. Zhou, "Deep learning in medical image analysis," in *Deep Learning in Medical Image Analysis: Challenges and Applications*. Cham: Springer International Publishing, 2020, pp. 3–21.

[68] X. Xu *et al.*, "A comprehensive review on synergy of multi-modal data and ai technologies in medical diagnosis," *Bioengineering*, vol. 11, no. 3, p. 3, Mar. 2024.

[69] S. A. Anisha, A. Sen, and C. Bain, "Evaluating the potential and pitfalls of ai-powered conversational agents as humanlike virtual health carers in the remote management of noncommunicable diseases: Scoping review," *Journal of Medical Internet Research*, vol. 26, no. 1, p. e56114, Jul. 2024.

[70] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[71] Y. Wang *et al.*, "A systematic review on affective computing: emotion models, databases, and recent advances," *Information Fusion*, vol. 83–84, pp. 19–52, Jul. 2022.

[72] K. Lesia Viktorivna, V. Andrii Oleksandrovych, K. Iryna Oleksandrivna, and K. Nadia Oleksandrivna, "Artificial intelligence in language learning: What are we afraid of," Online, 2022, accessed: May 29, 2025. [Online]. Available: https://eric.ed.gov/?id=EJ1363313

[73] L. Chen, P. Chen, and Z. Lin, "Artificial intelligence in education: A review," *IEEE Access*, vol. 8, pp. 75 264–75 278, 2020.

[74] S. Jabeen, X. Li, M. S. Amin, O. Bourahla, S. Li, and A. Jabbar, "A review on methods and applications in multimodal deep learning," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 19, no. 2s, pp. 76:1–76:41, Feb. 2023.

[75] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations & trends in multimodal machine learning: Principles, challenges, and open questions," *ACM Computing Surveys*, vol. 56, no. 10, pp. 264:1–264:42, Jun. 2024.

[76] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" *arXiv*, Apr. 2022.

[77] L. Zeng, Z. Shi, S. Xu, and D. Feng, "Safevanish: An improved data self-destruction for protecting data privacy," in *2010 IEEE Second International Conference on Cloud Computing Technology and Science*, Nov. 2010, pp. 521–528.

[78] E. Ntoutsi *et al.*, "Bias in data-driven artificial intelligence systems–an introductory survey," *WIREs Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1356, 2020.

[79] M. Cukurova, C. Kent, and R. Luckin, "Artificial intelligence and multimodal data in the service of human decision-making: A case study in debate tutoring," *British Journal of Educational Technology*, vol. 50, no. 6, pp. 3032–3046, 2019.

[80] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv*, Aug. 2017.

[81] G. Joshi, R. Walambe, and K. Kotecha, "A review on explainability in multimodal deep neural nets," *IEEE Access*, vol. 9, pp. 59 800–59 821, 2021.

[82] F. van Harmelen and A. ten Teije, "A boxology of design patterns for hybrid learning and reasoning systems," *Journal of Web Engineering*, vol. 18, no. 1–3, pp. 97–123, 2019.

[83] S. Malekmohamadi Faradonbe, F. Safi-Esfahani, and M. Karimian-kelishadrokhi, "A review on neural turing machine (ntm)," *SN Computer Science*, vol. 1, no. 6, p. 333, Oct. 2020.

[84] J. Huang, Z. Zhang, S. Zheng, F. Qin, and Y. Wang, "DISTMM: Accelerating distributed multimodal model training," in *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, 2024, pp. 1157–1171, accessed: May 29, 2025. [Online]. Available: https://www.usenix.org/conference/nsdi24/presentation/huang

[85] W. Zhao, T. Ma, X. Gong, B. Zhang, and D. Doermann, "A review of recent advances of binary neural networks for edge computing," *IEEE Journal on Miniaturization for Air and Space Systems*, vol. 2, no. 1, pp. 25–35, Mar. 2021.

[86] F. Zhuang *et al.*, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.

[87] M. Guimarães *et al.*, "Predicting model training time to optimize distributed machine learning applications," *Electronics*, vol. 12, no. 4, p. 4, Jan. 2023.

Received on June 12, 2025
Accepted on December 13, 2025