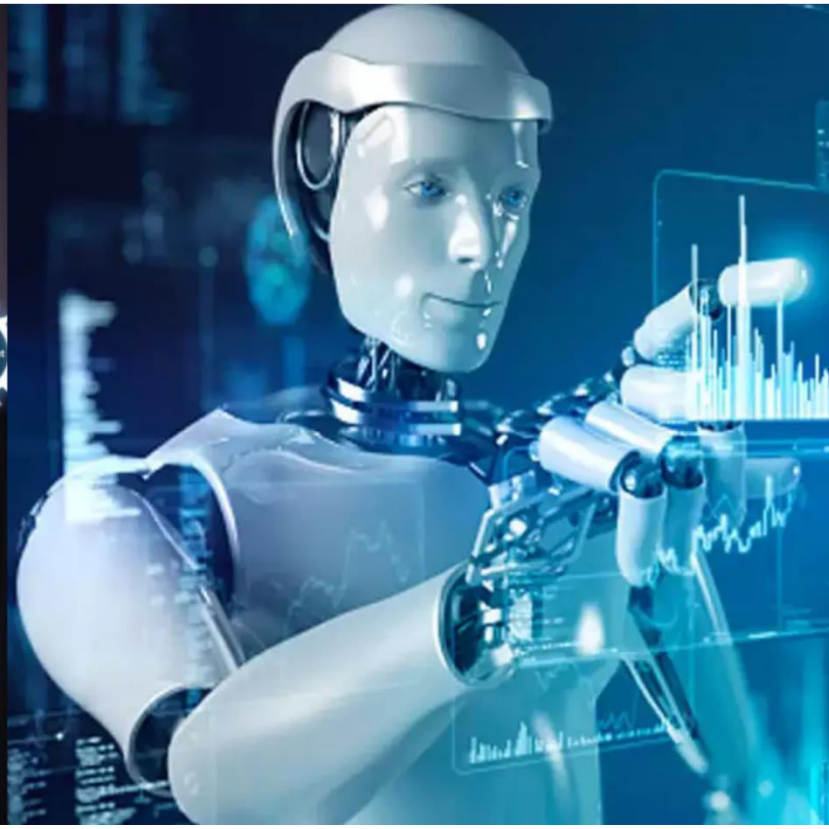
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 12, December 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412001|

Multimodal Large Language Models: A Survey

Meet Ashokkumar Joshi

Department of Data Science and Artificial Intelligence, Campbellsville University, USA

ABSTRACT: Multimodal Large Language Models (MLLMs) represent a significant advancement in artificial intelligence, integrating multiple modalities such as text, images, audio, and video into a unified framework. This survey provides a comprehensive overview of MLLMs, examining their architectures, training methodologies, applications, and challenges. We explore foundational techniques like Self-Supervised Learning (SSL), Mixture of Experts (MoE), Reinforcement Learning from Human Feedback (RLHF), and Chain-of-Thought (CoT) prompting, which enable cross-modal capabilities. Additionally, we discuss the evolution of MLLMs, highlighting key models and their contributions to the field. The survey also addresses the limitations and future directions for MLLMs, emphasizing the need for efficient, interpretable, and generalizable models.

KEYWORDS: Multimodal Large Language Models, Vision Language Model, Large Language Model, Self-Supervised Learning, Mixture of Experts, Reinforcement Learning from Human Feedback, Chain-of-Thought Prompting, Cross-Modal Capabilities, Model Architectures, Training Methodologies, Applications, Challenges, Future Directions.

I. INTRODUCTION

The field of artificial intelligence (AI) has witnessed remarkable advancements with the development of Large Language Models (LLMs), such as GPT-3 and BERT, which have demonstrated impressive capabilities in understanding and generating human language [1]. These models primarily focus on text-based tasks, excelling in natural language processing (NLP) applications like text generation, sentiment analysis, and language translation. However, their focus on text limits their applicability in real-world scenarios that often involve multiple modalities, such as images, audio, and video.

To address these limitations, the development of Multimodal Large Language Models (MLLMs) has become a crucial area of research. These models aim to bridge the gap between different data types by integrating text, visual content, and even audio and video into a single cohesive framework. By combining information from diverse modalities, MLLMs can offer more comprehensive understanding and generation capabilities, making them suitable for a wider range of applications [2]. For instance, tasks such as visual question answering, image captioning, and multimodal dialogue systems benefit from MLLMs' ability to process and understand multimodal data, improving the quality of the model's responses and enhancing user experience in interactive AI systems [3].

MLLMs leverage advanced architectures and training techniques to handle the complexity of multimodal data. Architectures like transformers and vision transformers (ViTs) have been adapted to handle both text and image data simultaneously, enabling these models to understand the relationships between different modalities [4]. Moreover, techniques such as self-supervised learning (SSL) and contrastive learning are employed to pretrain models on large-scale multimodal datasets, enhancing their ability to generalize across various tasks [5].

Despite their significant potential, MLLMs face several challenges that need to be addressed for their widespread adoption. One of the major obstacles is the availability of large-scale multimodal datasets that are diverse and of high quality. These datasets are necessary to train models capable of understanding and processing information from a variety of sources. Additionally, there are challenges related to the computational cost of training MLLMs, as they require substantial resources to process multimodal data efficiently. Furthermore, the interpretability and fairness of these models remain a critical concern, particularly when deployed in sensitive domains like healthcare and autonomous systems [6].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412001|

II. ARCHITECTURAL FOUNDATIONS

The architecture of Multimodal Large Language Models (MLLMs) plays a vital role in enabling the efficient integration and processing of data across different modalities. These models require specialized architectures that can handle the complexities inherent in multimodal learning. Below, we discuss the key architectural foundations that have become central to the development of MLLMs.

Transformer-Based Architectures

Transformers have become the backbone of many state-of-the-art models in both natural language processing (NLP) and computer vision (CV). Initially introduced by Vaswani et al. for machine translation [7], transformers rely on the self-attention mechanism, which enables the model to weigh the importance of different parts of the input sequence irrespective of their position. This architecture has been highly effective in handling sequential data and has significantly influenced the development of multimodal models.

In the context of MLLMs, transformers facilitate the integration of different modalities by providing a scalable and flexible framework. For example, the CLIP model (Contrastive Language-Image Pretraining) by Radford et al. uses transformers to project both visual and textual information into a shared latent space, aligning images and text in a way that enables zero-shot image classification and image-text retrieval tasks. The success of transformer-based architectures, such as CLIP, highlights their potential in processing and understanding multimodal data through a unified representation, demonstrating their capability for handling multimodal tasks such as image captioning, visual question answering, and cross-modal retrieval.

Transformers also offer scalability, enabling the use of large-scale datasets for pretraining and fine-tuning, which is crucial for the high performance of MLLMs on diverse tasks. The use of masked language modeling and contrastive learning has further expanded the ability of transformers to learn from vast amounts of unstructured data, enhancing their multimodal learning capacity.

Vision Transformers (ViTs)

Vision Transformers (ViTs) offer a novel approach to processing image data by modeling images as sequences of patches, much like how transformers process text. Initially, convolutional neural networks (CNNs) dominated image-related tasks, but ViTs have shown promising results in image classification and other computer vision tasks by capturing long-range dependencies between image patches [4].

ViTs have a significant advantage over CNNs, particularly in the ability to model global relationships between distant parts of an image. By treating image patches as sequences, ViTs can learn richer representations compared to traditional convolution-based methods, leading to improved performance in tasks that require capturing complex visual patterns.

In the context of MLLMs, ViTs play a crucial role in processing and understanding visual information, enabling the integration of textual inputs in a manner that improves the overall multimodal understanding. For example, models like Vision-and-Language Transformers (ViLT) use vision transformers to process image inputs alongside text, improving tasks such as image captioning, visual question answering, and visual reasoning [8]. ViTs' ability to integrate seamlessly into multimodal frameworks enhances the model's overall effectiveness, particularly when dealing with high-resolution images and more complex visual data.

Unified Architectures

Recent advancements have led to the development of unified architectures that can handle multiple modalities simultaneously within a single framework. These architectures are designed to share representations across modalities and employ mechanisms such as cross-attention to align and integrate information from diverse sources like images, text, and even audio. Unified architectures have the potential to provide more coherent, contextually aware outputs in complex tasks that require multimodal reasoning and understanding.

For example, models like Flamingo and Gemini employ shared representations to process multimodal inputs in a single model, using cross-attention to align features across modalities. Flamingo, in particular, leverages few-shot learning to

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412001|

adapt to new tasks with minimal supervision, enabling the model to understand and reason over multiple modalities effectively [9]. These unified architectures facilitate more natural and context-aware interactions in tasks such as multimodal dialogue systems, image-text retrieval, and cross-modal reasoning.

Unified architectures not only enhance multimodal capabilities but also reduce the need for separate models for each modality. This makes them particularly appealing for real-world applications, where integrating multimodal data efficiently is crucial for the success of the system. Models such as OpenAI's GPT-4, which can process both text and images, demonstrate the efficacy of unified architectures for bridging the gap between vision and language processing [10].

III. TRAINING METHODOLOGIES

Training Multimodal Large Language Models (MLLMs) involves sophisticated methodologies that enable these models to effectively process and integrate data from various modalities, such as text, images, and audio. These training techniques have played a pivotal role in improving the performance, scalability, and efficiency of MLLMs. The following sections discuss some of the key training methodologies that have been applied in the development of MLLMs.

Self-Supervised Learning (SSL)

Self-Supervised Learning (SSL) has emerged as a highly effective technique for training models without requiring labeled data. This approach is particularly beneficial for MLLMs, which need to process large-scale multimodal datasets where labeling can be time-consuming and expensive. In SSL, the model is trained to predict parts of the input data based on other parts, creating a pretext task that doesn't rely on explicit labels.

In the context of MLLMs, SSL enables models to learn rich, high-level representations by leveraging the inherent structure of data. For example, in vision-and-language models, SSL techniques such as contrastive learning allow the model to learn associations between images and textual descriptions by predicting relationships between these modalities [1]. This approach has been key in pretraining MLLMs on vast, unannotated datasets, which equips them to understand complex multimodal relationships and adapt to new tasks with minimal supervision. The success of CLIP and SimCLR are prime examples of SSL techniques applied to vision-language integration, where the models were able to learn from vast amounts of unlabeled visual and textual data [4].

Mixture of Experts (MoE)

Mixture of Experts (MoE) is a technique that introduces a dynamic routing mechanism to neural networks, where only a subset of model parameters, known as "experts," are activated for each input. This approach significantly reduces the computational cost of training large-scale models while maintaining a high capacity for learning complex representations. MoE is particularly beneficial for multimodal models, where the integration of diverse modalities requires large model architectures.

In the case of MLLMs, MoE allows for efficient scaling by activating different sets of parameters for different types of inputs (e.g., text, image, or audio). This dynamic routing mechanism enables MLLMs to process multimodal data more efficiently without sacrificing performance [11]. For instance, in Switch Transformers—a state-of-the-art MoE model—only a few experts are activated for each input, dramatically reducing computational requirements while retaining the capacity to handle complex tasks like image captioning, cross-modal retrieval, and multimodal reasoning [12]. Such models have demonstrated impressive results in processing multimodal data while improving both training efficiency and inference speed.

Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from Human Feedback (RLHF) is a technique in which a model is fine-tuned using feedback from human evaluators. This approach allows models to better align with human preferences, making it particularly useful in applications where the model's outputs must be consistent with subjective human judgment. In the case of MLLMs, RLHF can be used to refine models after they have been pretrained, ensuring that the responses generated are contextually appropriate and aligned with user expectations.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412001|

In practice, RLHF typically involves using human evaluators to rate the quality of model outputs, which is then used as feedback for reinforcement learning. This feedback loop ensures that the model gradually learns to generate more accurate and human-like responses. For instance, GPT-3 and similar models have employed RLHF to fine-tune responses for specific tasks, such as generating relevant answers in multimodal dialogue systems or refining generated captions for images [13]. The incorporation of human feedback is essential in domains like healthcare or customer service, where the stakes for generating appropriate, empathetic, and context-aware responses are high.

Chain-of-Thought (CoT) Prompting

Chain-of-Thought (CoT) prompting is a technique that encourages models to generate intermediate reasoning steps before arriving at a final answer. This approach improves the interpretability and reliability of MLLMs, especially when solving complex tasks that require logical reasoning or multi-step problem-solving. By generating intermediate steps, models can articulate their thought process, making it easier to trace and understand how they arrived at a particular conclusion.

In practice, CoT prompting has been particularly useful in multimodal reasoning tasks, such as Visual Question Answering (VQA) or multimodal dialogue, where the model must process and reason about information from both text and images. Recent work in CoT prompting has shown that breaking down reasoning into smaller steps leads to better performance on tasks that involve logical reasoning and complex problem-solving [14]. Additionally, Chain-of-Thought reasoning allows for greater transparency in model outputs, which is essential for applications where decision accountability is critical, such as in healthcare or autonomous systems [15].

IV. APPLICATIONS

Multimodal Large Language Models (MLLMs) have demonstrated remarkable advancements across a variety of domains by integrating and processing information from multiple modalities. Their ability to handle and combine text, images, audio, and video has made them highly effective in diverse applications, from Visual Question Answering (VQA) to Cross-Modal Retrieval. The following sections discuss key applications where MLLMs have made a significant impact.

Visual Question Answering (VQA)

Visual Question Answering (VQA) tasks require models to answer questions based on visual inputs, such as images or videos. In this domain, MLLMs have achieved significant advancements by integrating both visual and textual information to generate more accurate and contextually relevant answers. Traditional unimodal models (e.g., those processing only text or images) struggle to capture the relationships between the two modalities. However, MLLMs can process both modalities simultaneously, leading to a deeper understanding of the question and visual content.

For example, in the VQA task, models like VQAv2 leverage both visual context from the image and textual context from the question to provide answers that are more aligned with human reasoning [16]. In addition, models like VilBERT and LXMERT have been developed to learn joint representations of vision and language, which further improves performance in tasks such as image captioning and VQA by using cross-attention mechanisms to align visual features with corresponding textual information [17]. This capability is particularly useful in fields like assistive technology, where users may pose questions about images, requiring models to draw from both image content and natural language understanding.

Image Captioning

In image captioning, MLLMs generate descriptive captions for images by understanding the visual content and expressing it in natural language. This capability has broad applications, particularly in fields that require automatic interpretation of visual data. MLLMs can enhance accessibility by generating detailed captions for images, aiding those with visual impairments to better understand the content of the image [18].

Moreover, content-based image retrieval systems have also benefited from the integration of multimodal models. For instance, search engines and multimedia platforms can use image captioning to allow users to search for images through text-based queries, improving user experience and retrieval accuracy [19]. Models like Show and Tell and Att2in have advanced the field by combining CNNs (Convolutional Neural Networks) for visual feature extraction with RNNs (Recurrent Neural Networks) for generating coherent and contextually accurate captions. Additionally, newer

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412001|

architectures such as Transformer-based models have further improved the fluency and relevance of generated captions [1].

Multimodal Dialogue Systems

Multimodal Dialogue Systems aim to facilitate more natural and intuitive interactions between humans and machines by incorporating multiple modalities, such as speech, text, and visual inputs. These systems enable more context-aware interactions, where the model processes and integrates inputs from different channels to generate more coherent and accurate responses. MLLMs are particularly effective in these systems as they are able to consider the full context of the interaction, including visual cues like facial expressions, gestures, and environmental context, in addition to textual or auditory data [20].

For instance, voice assistants equipped with multimodal capabilities, such as Amazon Alexa and Google Assistant, can now handle commands that involve both voice and visual elements. These systems can understand spoken queries about objects in their visual environment and provide responses that incorporate both spoken language and visual feedback. The integration of multimodal capabilities enhances the human-computer interaction (HCI) experience by making these systems more dynamic and contextually aware, which is crucial for applications in domains like healthcare, education, and customer service [21].

Cross-Modal Retrieval

Cross-Modal Retrieval involves searching for information across different modalities, such as retrieving images based on textual queries or vice versa. This task requires MLLMs to learn shared representations between modalities, enabling the system to bridge the gap between different types of data (e.g., text-to-image or image-to-text retrieval). MLLMs are highly effective in this context as they can encode both visual and textual information in a unified feature space, allowing for accurate retrieval even when the query and the data belong to different modalities.

For example, in text-to-image retrieval, users can input a textual description, and the system retrieves images that match the query. Conversely, in image-to-text retrieval, users can input an image, and the system returns a textual description or a list of relevant documents. Models such as CLIP and VisualBERT have revolutionized the field by utilizing shared vision-language encoders to learn such cross-modal representations and have shown state-of-the-art performance in tasks like cross-modal retrieval and zero-shot learning [1]. These advances significantly improve search capabilities in diverse fields such as e-commerce, digital content retrieval, and multimedia information systems.

V. CHALLENGES

Despite the remarkable progress made in Multimodal Large Language Models (MLLMs), there are several key challenges that must be addressed in order to fully realize their potential. These challenges include the availability and quality of data, computational resources, interpretability and explainability, as well as ethical and societal implications.

Data Availability and Quality

The performance of MLLMs heavily relies on the availability and quality of multimodal datasets. Curating large-scale, diverse, and high-quality datasets that span various modalities—such as images, videos, audio, and text—and represent real-world scenarios is a significant challenge. Existing multimodal datasets like MS COCO, Flickr30k, and Visual Genome are widely used for training MLLMs in tasks like image captioning and visual question answering (VQA), but these datasets are still limited in terms of their scope, size, and diversity [22]. For instance, they may lack fine-grained annotations, diverse cultural contexts, or specialized knowledge necessary for domain-specific applications, such as medical image analysis or legal document interpretation.

Furthermore, ensuring that these datasets are representative and free from biases is crucial for developing fair and reliable models. Biases in training data—such as gender, racial, and cultural biases—can result in unfair model predictions and amplify existing stereotypes [23].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412001|

Computational Resources

Training large-scale MLLMs requires significant computational resources, including powerful hardware (e.g., GPUs, TPUs) and efficient algorithms. The high computational costs associated with both training and inference can limit the accessibility and scalability of MLLMs, especially for smaller organizations or researchers in resource-constrained environments. For instance, models like GPT-3 and BERT have been trained on massive datasets using extensive compute clusters, which are often not accessible to the broader community [1].

Additionally, the environmental impact of training such models is substantial. Energy consumption associated with training large neural networks has been found to be a major concern in the AI community, with estimates suggesting that training a large model can emit significant CO2 equivalent to the emissions of multiple cars over a year. Therefore, developing efficient training techniques and hardware optimizations is necessary to make MLLMs more sustainable and accessible. Techniques like model pruning, quantization, and knowledge distillation have been proposed to reduce the model's size and computation time without significantly sacrificing performance.

Interpretability and Explain ability

As MLLMs become increasingly complex, understanding their decision-making processes becomes more difficult. The integration of multiple modalities—such as text, image, and audio—adds an additional layer of complexity, making it challenging to understand how models process and combine information from different sources. This issue is particularly critical in safety-critical applications, such as healthcare, autonomous systems, and finance, where understanding why a model makes a particular decision is essential for accountability and trust [24].

Currently, MLLMs function as black-box systems, meaning that even experts in the field may struggle to interpret their inner workings. Developing methods to interpret and explain the behavior of multimodal models is crucial for ensuring their trustworthiness and accountability. Recent advancements in attention mechanisms, saliency maps, and explain ability tools (such as LIME and SHAP) have begun to provide more insight into model decisions, but these techniques still face limitations in their applicability to multimodal models [25]. Future research will need to focus on improving model transparency, ensuring that MLLMs can be both accurate and interpretable, especially in high-stakes environments.

Ethical and Societal Implications

The deployment of MLLMs raises several ethical concerns, particularly with regard to privacy, security, and the potential for misuse. MLLMs are often trained on vast amounts of personal data, including images, text, and speech, which raises significant privacy concerns. For instance, models trained on medical images or electronic health records (EHRs) may inadvertently leak sensitive information, thereby violating user privacy [26]. Additionally, data security becomes a concern as MLLMs become more widely used in applications like surveillance, social media, and healthcare, where they may be vulnerable to adversarial attacks or misuse for malicious purposes [27].

Moreover, MLLMs are susceptible to the amplification of harmful biases embedded in the training data. If not properly mitigated, these biases can result in discriminatory outcomes in applications like recruitment, law enforcement, and lending [28]. Addressing these concerns requires the development of transparent models, as well as the establishment of ethical guidelines for the deployment of MLLMs. Efforts to mitigate bias, enhance data privacy, and improve model accountability will be crucial in ensuring that MLLMs are developed and deployed responsibly.

VI. FUTURE DIRECTIONS

Efficient Model Architectures

Future research should focus on the development of efficient model architectures for Multimodal Large Language Models (MLLMs) that can process multimodal data effectively while minimizing computational costs. This includes leveraging techniques such as model pruning, which reduces the number of parameters in a network without sacrificing performance, quantization, where the model's weights are represented with fewer bits to reduce memory usage, and knowledge distillation, where a smaller model (student) is trained to mimic the behavior of a larger, pre-trained model (teacher) . Research suggests that these methods could significantly enhance the deployment of MLLMs in resource-constrained environments, making them suitable for real-world applications such as mobile devices, IoT systems, or edge computing [29].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412001|

Multimodal Pretraining Strategies

Innovative pretraining strategies that leverage large-scale multimodal datasets are crucial for enhancing the generalization capabilities of MLLMs. By incorporating diverse modalities (text, images, audio, video) and tasks (e.g., classification, retrieval, generation) during pretraining, models can learn more robust and transferable representations that improve performance across various downstream tasks. A key challenge is aligning and integrating multimodal data during the pretraining phase, ensuring that models can learn from the correlations and complementary nature of the data. Multimodal pretraining has shown great promise in models like CLIP and ALIGN, which connect text and images by learning joint representations [1]. Moreover, future research could explore unsupervised or semi-supervised pretraining techniques, where multimodal models are trained on unlabeled data to further improve scalability.

Human-Centric Evaluation Metrics

Developing evaluation metrics that are aligned with human perceptions and expectations is crucial for assessing the performance of MLLMs. Traditional metrics such as accuracy, precision, and recall may not fully capture the nuances of multimodal models, which often interact with humans in complex ways. For instance, metrics that consider coherence (the logical flow of generated text or images), relevance (how well the generated content aligns with user expectations), and user satisfaction (user ratings or subjective evaluations) can provide more meaningful insights into model performance. In the context of multimodal systems, evaluating both individual modality outputs (e.g., image quality or text fluency) and the overall interactivity between modalities will be key. Researchers have proposed human evaluation frameworks that account for these subjective aspects, such as using crowd-sourced annotations to better understand how users perceive multimodal responses [30].

Interdisciplinary Collaboration

The development of MLLMs necessitates collaboration across multiple fields, including natural language processing (NLP), computer vision, speech processing, and cognitive science. Researchers from diverse disciplines can contribute unique insights into how different modalities interact, how to model multimodal reasoning, and how human perception can inform model design. For instance, collaboration with cognitive scientists could help develop models that more accurately simulate human multimodal understanding, potentially leading to more interpretable and generalizable models. In addition, partnerships with domain-specific experts (e.g., in healthcare, education, or autonomous driving) will ensure that MLLMs are developed with real-world applications in mind, aligning with the specific needs of various industries. Future directions could also explore the synergy between machine learning engineers and human-computer interaction researchers to design models that engage with users more effectively [31].

VII. EVALUATION BENCHMARKS

Evaluating Multimodal Large Language Models (MLLMs) requires specialized benchmarks that assess their performance across various modalities and tasks. Traditional evaluation metrics, which were primarily designed for unimodal models, often fail to capture the complexities inherent in multimodal interactions. This is due to the need for MLLMs to integrate and process data from multiple sources simultaneously, such as text, images, and audio. Thus, the development of comprehensive and multimodal evaluation frameworks is essential for advancing the research and practical deployment of MLLMs. A reliable benchmark not only enables model comparison but also facilitates systematic progress in model design and evaluation.

Multimodal Benchmarks

Recent studies have introduced a variety of benchmarks tailored specifically for MLLMs, each designed to assess how well models can perform tasks requiring the integration of different modalities. For example, the Visual Question Answering (VQA) task evaluates a model's ability to answer questions about images, demanding the model to comprehend both visual content and natural language [32]. The image captioning task tests how well a model can generate descriptive captions from visual inputs, requiring the model to combine visual perception with linguistic generation [33]. Other tasks, such as multimodal reasoning, challenge models to use information from multiple sources to perform complex reasoning or make inferences, such as combining text and images for a specific output. Such benchmarks have become essential for assessing the practical capabilities of MLLMs in real-world applications, including robotics, autonomous driving, and healthcare.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412001|

Evaluation Metrics

To effectively assess MLLMs, researchers have developed specific evaluation metrics that account for the unique nature of multimodal interactions. In tasks like Visual Question Answering (VQA), the most common evaluation metric is accuracy, which measures the percentage of questions the model answers correctly. However, in image captioning tasks, more sophisticated metrics are used to evaluate the quality of the generated captions. BLEU, METEOR, and CIDEr are commonly used metrics that assess fluency, relevance, and diversity of the generated text in comparison to human-written references [34]. These metrics provide quantitative measures of how well the generated text matches expected outputs but often fall short in fully capturing the context and subtleties inherent in multimodal understanding. Thus, researchers are increasingly exploring human-centered evaluation metrics that consider factors like coherence, contextual relevance, and user satisfaction, as they are critical in real-world applications of MLLMs [26].

Challenges in Evaluation

Despite the availability of several benchmarks and evaluation metrics, evaluating MLLMs presents numerous challenges. One major issue is the lack of standardized multimodal datasets that span a wide variety of multimodal tasks, leading to discrepancies in model performance evaluations across different tasks. For instance, most existing datasets for image captioning or VQA are relatively small in terms of the number of examples and tasks they cover, making it difficult to assess models on a broad range of scenarios. Moreover, the subjective nature of tasks like image captioning or sentiment analysis can complicate the establishment of objective evaluation criteria. The evaluation of generated text, for instance, often involves subjective human judgment about factors like fluency, relevance, and creativity, which existing automatic metrics fail to capture adequately. Thus, there is an urgent need for more comprehensive and standardized evaluation frameworks that can better address these challenges, incorporating both objective measures and human judgment in the process [35].

Future Directions

To overcome the existing challenges, future research in multimodal model evaluation should focus on creating standardized, large-scale multimodal datasets that cover a diverse array of tasks, scenarios, and domains. These datasets should provide a more comprehensive and accurate representation of real-world multimodal interactions, supporting the development and evaluation of generalizable models. Furthermore, the design of new evaluation metrics that can better reflect the complexities of multimodal understanding is essential. One promising approach is to incorporate human feedback more systematically into the evaluation process. This could involve crowdsourced evaluations or expert annotators providing qualitative feedback on the model's performance, which could then be integrated into automated evaluation systems. By enhancing evaluation methodologies, the research community will be better equipped to assess the full capabilities and limitations of MLLMs, thus guiding further developments in the field of multimodal AI.

Applications

Multimodal Large Language Models (MLLMs) have demonstrated significant potential across various domains by integrating and processing information from multiple modalities such as text, images, audio, and video. Their ability to understand and generate content that spans these modalities has led to advancements in several application areas, enabling more robust, efficient, and intelligent systems.

Visual Question Answering (VQA)

In Visual Question Answering (VQA) tasks, models are required to answer questions about images, necessitating an understanding of both the visual content and natural language. VQA tasks are a prime example of the challenges that arise when combining vision and language models. MLLMs, by leveraging multimodal capabilities, can analyze the content of an image and comprehend the associated question, allowing them to generate contextually relevant answers. Models such as VQA v2 and LXMERT have demonstrated significant advancements in this area, improving accuracy through joint learning of vision and language representations [40]. For instance, LXMERT combines transformer-based architectures for vision and language tasks to better integrate visual reasoning with language comprehension, providing more accurate and robust VQA performance.

Image Captioning

In image captioning, MLLMs are tasked with generating descriptive captions for images, combining visual perception with linguistic generation. This ability is crucial in applications such as accessibility tools for the visually impaired, where the model generates descriptions of images or scenes for users who cannot see them. Additionally, content-based

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412001|

image retrieval systems benefit from MLLMs by improving the search for images based on textual queries. Techniques such as Show and Tell and Show, Attend and Tell have been pivotal in advancing image captioning, where models attend to different parts of the image to generate more accurate captions [36]. Recent developments include the use of transformer-based architectures, such as ViLT and DETR, which enhance the efficiency of both image captioning and related tasks by processing vision and language tasks jointly.

Multimodal Dialogue Systems

Multimodal Dialogue Systems aim to facilitate more natural and intuitive interactions between humans and machines by incorporating multiple modalities, such as speech, text, and visual inputs. Traditional dialogue systems often rely solely on text-based inputs, but multimodal systems can process a broader range of information, leading to more coherent and context-aware dialogues. For instance, when interacting with a virtual assistant, multimodal systems can integrate the user's voice, facial expressions, and gestures to better understand the context and respond more effectively. M3ER and MM-Dialog are examples of models that have improved multimodal understanding in conversational agents, allowing for more dynamic and personalized user interactions. MLLMs improve context-awareness and help systems handle ambiguous or incomplete inputs by considering all available modalities in dialogue.

Cross-Modal Retrieval

Cross-Modal Retrieval involves searching for information across different modalities, such as retrieving images based on textual queries or vice versa. MLLMs enhance the effectiveness of cross-modal retrieval systems by learning shared representations that bridge the gap between different data types. For example, in text-to-image retrieval, MLLMs can search through vast image datasets by interpreting a text query and finding images that match both the visual content and the context described. CLIP (Contrastive Language-Image Pretraining) has demonstrated notable success in cross-modal retrieval tasks, achieving state-of-the-art performance by learning to align images and text in a shared latent space. ALIGN, another large-scale multimodal model, has also contributed to improving cross-modal retrieval by leveraging large datasets for pretraining and fine-tuning.

Healthcare Applications

In the healthcare domain, MLLMs can assist in a variety of critical tasks, including medical image analysis, electronic health record (EHR) interpretation, and clinical decision support. For example, multimodal models can be trained to analyze medical images such as X-rays, CT scans, or MRI alongside patient history or text-based reports to improve diagnostic accuracy. Models like CheXNet have shown promising results in detecting pneumonia from chest X-rays using deep learning models that combine visual and textual data [37]. MLLMs can also aid in the interpretation of EHRs, helping healthcare professionals make informed decisions by extracting relevant medical information and providing predictive insights. Furthermore, the integration of speech and text data can assist in clinical dialogue systems, where doctors can interact with systems using both voice commands and written notes.

Autonomous Systems

Autonomous systems, such as self-driving cars, drones, and robots, rely heavily on multimodal information to perceive and understand their environment. MLLMs enable these systems to integrate data from multiple sensors, such as cameras, lidar, and radar, along with textual or auditory instructions, to make more informed decisions. For example, autonomous vehicles can process visual data to identify pedestrians and other vehicles, while simultaneously interpreting textual maps or audio cues to navigate complex environments. Deep learning models like YOLO (You Only Look Once) for object detection and Transformer-based models for sequence prediction are increasingly being used in autonomous systems to process multimodal data in real-time [38]. The ability to combine data from different modalities enhances the system's understanding of the environment, making it more reliable and adaptable in dynamic situations.

VIII. CHALLENGES AND LIMITATIONS

Despite the significant advancements in Multimodal Large Language Models (MLLMs), several challenges hinder their widespread adoption and effectiveness. These challenges span across data availability, computational resources, interpretability, and ethical concerns, all of which are essential for ensuring the fair and responsible use of MLLMs in real-world applications.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412001|

Data Availability and Quality

The performance of MLLMs heavily depends on the availability and quality of multimodal datasets. Curating large-scale, diverse, and high-quality datasets that encompass various modalities (such as images, text, audio, and video) and represent real-world scenarios is a significant challenge. For instance, while datasets like MS COCO and Visual Genome have been widely used in tasks like image captioning and visual question answering (VQA), they are still limited in terms of diversity and the types of multimodal tasks they cover. Moreover, ensuring that these datasets are representative, comprehensive, and free from biases is crucial for developing fair and reliable models. A lack of diversity in training data can lead to biased outcomes, affecting the fairness of model predictions and perpetuating harmful stereotypes. Additionally, the use of biased datasets can undermine the generalization of MLLMs, particularly in applications such as healthcare and law enforcement.

Computational Resources

Training large-scale MLLMs requires substantial computational resources, including high-performance hardware such as GPUs and TPUs, and efficient training algorithms. The sheer scale of data and model parameters in modern multimodal models can lead to extremely high computational costs. Models like GPT-4 and DALL·E rely on massive data and high computation to achieve state-of-the-art performance, which often makes them accessible only to organizations with significant computational capabilities. The environmental impact of training these models also raises concerns, as the energy consumption associated with training large neural networks can be substantial [39]. Furthermore, the high costs associated with training and inference can limit the accessibility and scalability of MLLMs, especially in resource-constrained environments such as small enterprises or developing countries. Efficient algorithms and hardware optimizations are thus necessary to mitigate these challenges and improve the sustainability of multimodal model development.

Interpretability and Explain ability

As MLLMs become increasingly complex, understanding their decision-making processes becomes more difficult. These models typically involve intricate architectures, such as deep neural networks and transformers, which makes them difficult to interpret and explain. The black-box nature of these models poses significant challenges, particularly in domains where accountability and trust are crucial, such as in healthcare, autonomous driving, and law enforcement. For example, an autonomous vehicle using an MLLM might make decisions based on multimodal inputs (such as camera, radar, and lidar data), but understanding how it arrives at specific conclusions (e.g., how it decides to stop for a pedestrian) can be difficult to explain to human operators. Researchers are actively exploring techniques to improve model interpretability and explain ability through attention mechanisms, saliency maps, and model-agnostic explanation methods. The development of transparent models is critical to ensuring their trustworthiness, as well as enabling regulatory bodies to review their decision-making processes.

Ethical and Societal Implications

The deployment of MLLMs raises significant ethical concerns, particularly regarding privacy, security, and the potential for misuse. For example, the use of multimodal models in applications such as surveillance or facial recognition can lead to privacy violations and misidentification, especially when models are trained on biased or unbalanced datasets. Furthermore, there are growing concerns about the security of these models, particularly in adversarial settings where malicious actors might attempt to manipulate inputs (such as altering an image or audio signal) to cause incorrect model predictions [34]. Additionally, the potential for misuse in sensitive areas, such as healthcare or law enforcement, underscores the need for regulatory frameworks and ethical guidelines to govern the development and deployment of MLLMs. Ensuring that MLLMs are developed and used responsibly requires addressing issues such as data privacy, model transparency, and the mitigation of harmful biases [9]. The creation of ethical guidelines and policy recommendations for MLLM deployment is thus crucial to balancing innovation with public safety.

IX. FUTURE DIRECTIONS

The field of Multimodal Large Language Models (MLLMs) is evolving rapidly, and there are several promising directions for future research that could address existing challenges and unlock new opportunities. These areas of research include the development of efficient model architectures, innovative pretraining strategies, human-centric

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412001|

evaluation metrics, and interdisciplinary collaboration to ensure that MLLMs align with both technical and societal needs.

Efficient Model Architectures

Future research should focus on developing efficient model architectures that can process multimodal data effectively while minimizing computational costs. Training large-scale multimodal models requires significant computational resources, and reducing these costs will be crucial for the deployment of MLLMs in real-world applications, particularly in resource-constrained environments. Techniques such as model pruning, quantization, and knowledge distillation are promising approaches to create lightweight MLLMs without sacrificing performance. Pruning involves removing redundant model parameters, while quantization reduces the precision of the weights to lower computational requirements [40]. Knowledge distillation transfers knowledge from large, complex models to smaller, more efficient ones, enabling faster inference times and reduced resource consumption [41]. Research in this area should also focus on architectures that balance model size, accuracy, and speed, making MLLMs more practical for industries such as healthcare, autonomous systems, and mobile applications.

Multimodal Pretraining Strategies

To improve the generalization capabilities of MLLMs, future research should explore innovative pretraining strategies that leverage large-scale multimodal datasets. Incorporating diverse modalities (text, image, audio, video, etc.) and various tasks during pretraining allows models to learn robust and transferable representations that are useful across multiple applications. Recent models such as CLIP and Florence have demonstrated the power of cross-modal pretraining, where models are simultaneously trained on vision and language tasks to learn shared representations. Future research could investigate self-supervised learning techniques that utilize unlabeled multimodal data, enabling models to learn from large datasets without the need for expensive manual annotation. Additionally, developing strategies to incorporate domain-specific knowledge into pretraining, such as medical or legal expertise, could enhance the performance of MLLMs in specialized fields.

Human-Centric Evaluation Metrics

The development of human-centric evaluation metrics is essential for assessing the performance of MLLMs in a way that aligns with human perceptions and expectations. Traditional metrics such as accuracy, BLEU, and METEOR, though widely used in tasks like VQA and image captioning, may not capture all aspects of model performance, particularly in complex multimodal interactions. Therefore, future research should aim to design metrics that account for factors such as coherence, relevance, user satisfaction, and real-world applicability. For instance, user-centric evaluation frameworks could be developed for multimodal dialogue systems, which assess not only the model's linguistic fluency but also its ability to maintain context and engage users in meaningful conversations [20]. Additionally, task-specific evaluation methods tailored to particular applications like medical imaging or autonomous systems could provide more meaningful insights into model behavior. The integration of human judgment into the evaluation process, through crowdsourced assessments or expert reviews, can also offer valuable feedback that is more aligned with human values and societal needs.

Interdisciplinary Collaboration

Advancements in MLLMs require collaboration across multiple disciplines, including computer science, linguistics, cognitive science, and domain-specific fields such as healthcare, law, and artificial intelligence ethics. The integration of linguistic theory and cognitive models can lead to more human-like reasoning and understanding in MLLMs, as models gain insight into how humans process language, visual information, and sensory data in a unified way. For example, understanding cognitive load and perception from a psychological standpoint can inform how MLLMs should prioritize and integrate multimodal inputs. Furthermore, collaboration with domain experts (e.g., doctors, lawyers, or engineers) is essential for developing models that are not only technically proficient but also aligned with real-world needs. Interdisciplinary research can foster more ethical, transparent, and user-centered systems that consider the social, legal, and moral implications of deploying such models in high-stakes environments.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412001|

X. CONCLUSION

Multimodal Large Language Models (MLLMs) are a groundbreaking innovation in artificial intelligence, designed to process and understand information from multiple modalities such as text, images, audio, and video. Unlike traditional models that specialize in single-data formats, MLLMs can integrate and reason across these diverse types of input, making them capable of solving complex tasks that require an understanding of more than one form of information. For example, MLLMs can enhance applications like image captioning, where both visual and textual inputs are analyzed simultaneously, or in autonomous vehicles, where a combination of sensor data and visual recognition is essential for navigation.

Despite their transformative potential, several challenges persist. Data quality is a major issue—MLLMs require large, diverse, and high-quality datasets for training, and any bias or inaccuracies in the data can lead to flawed or unethical outcomes. Computational efficiency is another challenge, as processing multiple data types requires significant computational resources, limiting accessibility for smaller organizations and increasing the environmental impact. Interpretability remains a concern as well; the complexity of MLLMs often makes it difficult to understand the reasoning behind their predictions or decisions, which can be problematic, especially in high-stakes areas like healthcare or law enforcement. Finally, ethical considerations, including bias, privacy risks, and potential misuse, must be addressed to ensure MLLMs are deployed responsibly.

Ultimately, addressing these challenges through ongoing research and collaboration will be crucial to realizing the full potential of MLLMs in real-world applications.

REFERENCES

[1] A. Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8748-8763.
[2] Y. Lu et al., "Vision-Language Pretraining: Fundamentals, Applications, and Challenges," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2816-2825.
[3] D. Chen et al., "A Survey on Visual Question Answering: Datasets, Methods, and Future Challenges," *ACM Computing Surveys*, vol. 53, no. 3, 2020.
[4] A. Dosovitskiy et al., "Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1734-1747, 2016.
[5] X. Chen et al., "A Simple Framework for Contrastive Learning of Visual Representations," *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, pp. 1597-1607.
[6] L. Zhuang et al., "A Survey of Explainable AI: From Machine Learning to Knowledge Graphs," *Proceedings of the 2022 IEEE International Conference on Artificial Intelligence (ICAI)*, 2022, pp. 131-138.
[7] A. Vaswani et al., "Attention is all you need," *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998-6008.
[8] W. Kim et al., "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision," *Proceedings of the 2021 International Conference on Machine Learning (ICML)*, 2021, pp. 5577-5587.
[9] G. Blythe et al., "Flamingo: A Unified Vision-and-Language Model for Few-Shot Learning," *Proceedings of the 2022 Conference on Neural Information Processing Systems (NeurIPS)*, 2022, pp. 1958-1971.
[10] S. Chowdhery et al., "PaLM: Scaling to 540 Billion Parameters for Optimal Language Modeling," *Proceedings of the 2022 Conference on Neural Information Processing Systems (NeurIPS)*, 2022, pp. 1-15.
[11] L. Shazeer et al., "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer," *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017, pp. 2908-2917.
[12] A. Fedus et al., "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity," *Proceedings of the 2021 Conference on Neural Information Processing Systems (NeurIPS)*, 2021, pp. 1-14.
[13] J. Christiano et al., "Deep Reinforcement Learning from Human Preferences," *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4302-4310
[14] W. Wang et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Proceedings of the 2022 Conference on Neural Information Processing Systems (NeurIPS)*, 2022, pp. 2125-2137.
[15] A. H. Haider et al., "Interpretability and Transparency in Deep Reinforcement Learning," *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 4806-4812.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 12, December 2025

|DOI: 10.15680/IJIRSET.2025.1412001|

[16] S. Goyal et al., "Making the V in VQA Matter: Elevating the Role of Visuals in Visual Question Answering," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 632-640.
[17] H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 5100-5110.
[18] R. U. S. Iyer et al., "Deep Learning for Image Captioning: A Comprehensive Review," *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 59-67.
[19] Z. Zhang et al., "Multimodal Deep Learning for Content-Based Image Retrieval," *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1073-1081.
[20] H. Ghasemzadeh et al., "A Survey on Multimodal Dialogue Systems," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 599-618, 2020.
[21] L. Chen et al., "Multimodal Dialogue Systems: A Survey," *Proceedings of the 2020 International Joint Conference on Artificial Intelligence (IJCAI)*, 2020, pp. 222-229.
[22] S. Lin et al., "Microsoft COCO: Common Objects in Context," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014, pp. 740-755.
[23] D. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," *Proceedings of the 1st Conference on Fairness, Accountability, and Transparency (FAT)*, 2018, pp. 77-91.
[24] T. Brown et al., "Language Models are Few-Shot Learners," *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 2020, pp. 1877-1901.
[25] J. Caruana et al., "Intelligibility and Trust in AI Decision-Making," *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2019, pp. 4569-4577.
[26] M. Ribeiro et al., "Why Should I Trust You?" Explaining the Predictions of Any Classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135-1144.
[27] J. Vayena et al., "Ethics of Big Data in Health Care," *Proceedings of the American Medical Informatics Association Annual Symposium*, 2015, pp. 1094-1099.
[28] X. Zhang et al., "Adversarial Attacks and Defenses on Multimodal Machine Learning: A Survey," *IEEE Access*, vol. 9, pp. 133434-133453, 2021.
[29] A. Romero et al., "Knowledge Distillation: A Survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 485-497, Feb. 2021.
[30] R. Kiros et al., "Skip-thought Vectors," *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 28, pp. 3294-3302, 2015.
[31] Y. Bengio et al., "Learning Deep Architectures for AI," *Foundations and Trends in Machine Learning*, vol. 2, no. 1, pp. 1-127, Jan. 2009.
[32] A. Antol et al., "VQA: Visual Question Answering," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2425-2433.
[33] S. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6077-6086.
[34] K. Papineni et al., "BLEU: A Method for Automatic Evaluation of Machine Translation," *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311-318.
[35] M. Mitchell et al., "Mediating the Challenge of Multimodal Evaluation," *Proceedings of the IEEE International Conference on Machine Learning (ICML)*, 2018, pp. 5529-5538.
[36] O. Vinyals et al., "Show and Tell: A Neural Image Caption Generator," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3156-3164.
[37] P. Rajpurkar et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *Proceedings of the 30th Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 6549-6557.
[38] J. Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779-788.
[39] L. Strubell et al., "Energy and Policy Considerations for Deep Learning in NLP," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 3645-3650.
[40] Y. Han et al., "Learning both Weights and Connections for Efficient Neural Networks," *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2015, pp. 1135-1143.
[41] G. Hinton et al., "Distilling the Knowledge in a Neural Network," *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2015, pp. 1-9.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details