Multimodal Large Language Models in Healthcare: A Survey

Abstract

This survey paper explores the transformative potential of Multimodal Large Language Models (MLLMs) in healthcare, highlighting their ability to integrate diverse data modalities, such as text, images, and genomic information, to enhance clinical applications. MLLMs have shown significant advancements in diagnostic accuracy, personalized treatment planning, and overall patient care by synthesizing structured and unstructured data, thus facilitating more informed clinical decision-making. Despite their potential, challenges such as data acquisition, computational constraints, model limitations, and ethical considerations hinder their deployment in healthcare settings. Addressing these challenges is crucial for ensuring the reliability and ethical integration of MLLMs into clinical practice. The survey emphasizes the importance of developing comprehensive benchmarks and datasets, advancing model architecture, and training techniques to overcome these challenges. Furthermore, the integration of MLLMs with emerging technologies and fostering collaborative research efforts are pivotal for advancing medical AI. Collaborative initiatives among healthcare professionals, data scientists, and policymakers can drive innovative solutions, expanding the applicability of MLLMs across diverse healthcare settings. By prioritizing clinical validation and real-world application, MLLMs can be effectively integrated into healthcare systems, ultimately transforming healthcare delivery and improving patient care. The survey concludes with a call to action for continued research and interdisciplinary collaboration to fully realize the potential of MLLMs in revolutionizing healthcare.

1 Introduction

1.1 Significance and Potential Impact

The integration of Multimodal Large Language Models (MLLMs) in healthcare marks a significant advancement in artificial intelligence, with the potential to transform clinical practices and patient management. MLLMs synthesize structured and unstructured data from diverse modalities—text, images, and audio—enhancing diagnostic accuracy and supporting personalized treatment plans. By leveraging these varied data streams, MLLMs provide comprehensive insights into patient health, improving clinical decision-making and patient engagement [1, 2, 3]. Their capability for comprehensive reasoning across modalities is crucial for addressing complex healthcare challenges.

Remarkable advancements in domain-specific visual tasks, such as visual question answering, are vital for accurate medical diagnosis and treatment planning [4]. MLLMs also enhance the interpretation of complex medical images by understanding anatomical regions within scans, thereby improving clinical decision-making processes [5]. Furthermore, progress in vision-language tasks underscores their transformative impact on healthcare delivery [6], facilitating secure data management and effective data sharing, which are critical for advancing healthcare applications [7]. The enhancement of multimodal representation in item-to-item recommendations using MLLMs indicates their potential to improve recommendation accuracy in personalized healthcare solutions [8].

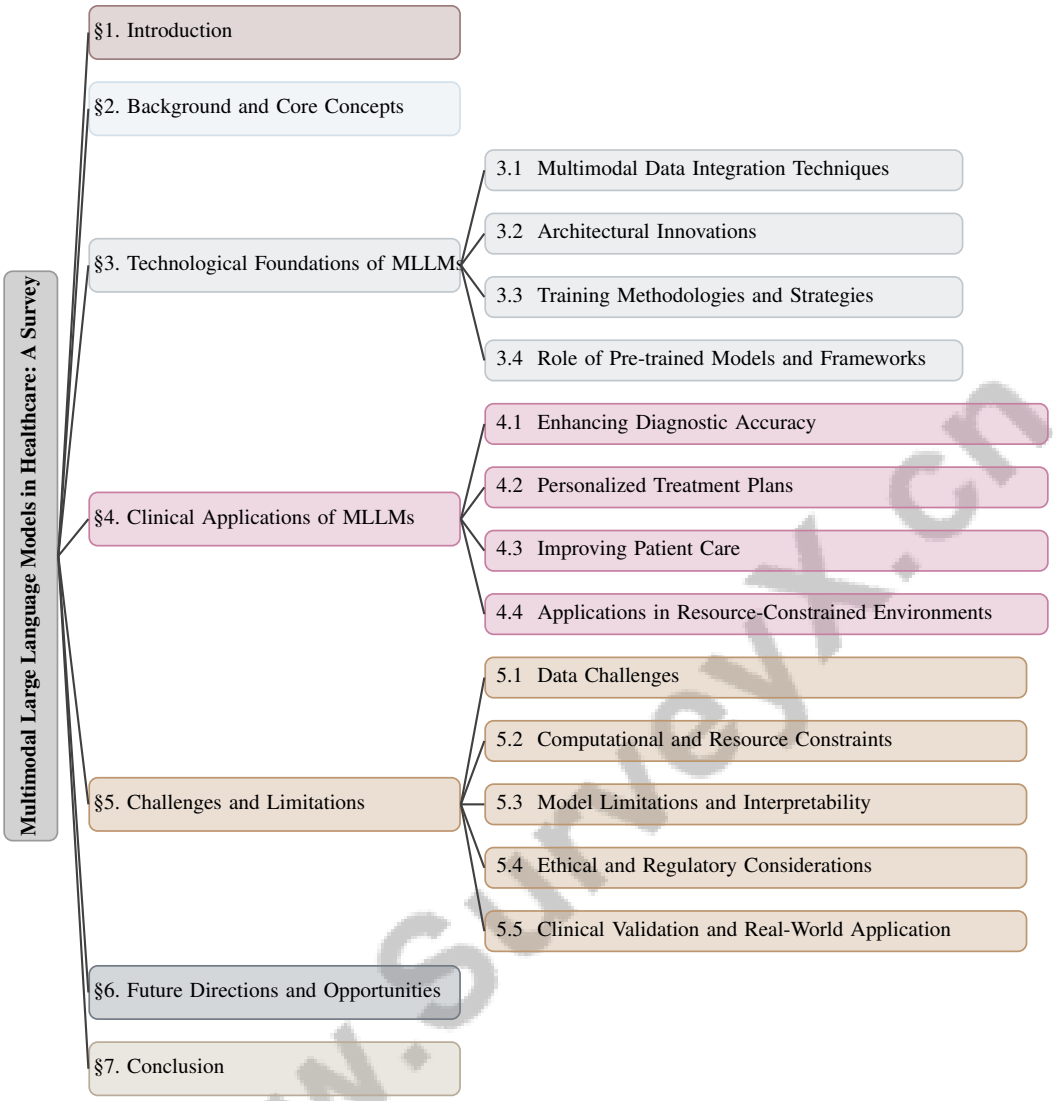


Figure 1: chapter structure

However, the increasing model size and computational complexity of MLLMs present challenges, particularly in resource-constrained environments [9]. Developing data-efficient and compute-efficient multimodal models is essential to realize their transformative potential across diverse healthcare settings [10]. Continued refinement and deployment of MLLMs are crucial to overcoming existing challenges and fully harnessing their capabilities in clinical environments.

1.2 Scope and Objectives of the Survey

This survey provides a comprehensive examination of Multimodal Large Language Models (MLLMs) within the healthcare sector, emphasizing their transformative potential across various clinical applications. It addresses the preprocessing, modeling, and evaluation stages of machine learning as applied to healthcare data, tackling the complexities and volume of such data [11]. The performance of both proprietary models like GPT-4 and Gemini, as well as six open-source MLLMs across modalities including text, code, image, and video, is evaluated [12].

Key areas of focus include the architecture, pre-training objectives, multilingual data construction, and evaluation methodologies of MLLMs, alongside their real-world applications [13]. The survey delves into various reasoning tasks that MLLMs can perform, such as visual question answering and multimodal dialogue, emphasizing high-level categorizations and evaluations [14].

The survey evaluates the integration of healthcare foundation models (HFMs) within clinical settings, concentrating on their direct medical applications and benefits. By excluding non-medical applications, it provides a focused analysis of HFMs' capabilities, challenges, and transformative potential in enhancing healthcare services, including clinical language processing, medical image analysis, and omics research, while addressing unique deployment obstacles [15, 16]. The utilization of Transformers for analyzing various healthcare data types, including medical imaging, electronic health records (EHRs), and biomolecular sequences, ensures a concentrated discourse on healthcare-specific applications.

Additionally, the survey explores AI applications in healthcare, covering diagnostics, personalized treatment, telehealth, and administrative efficiencies. It thoroughly examines the integration of AI in healthcare, focusing on disease diagnosis, treatment recommendations, and patient engagement, while addressing critical ethical considerations such as data privacy, informed consent, and algorithmic biases. The need for robust regulatory frameworks to ensure responsible implementation and accountability in AI technologies within clinical practice is underscored [17, 18].

The survey systematically reviews the applications of MLLMs across diverse tasks, identifying existing knowledge gaps to provide insights that facilitate the development of robust AI-driven solutions for healthcare professionals. This approach seeks to enhance the reliability of medical AI applications and bridge the gap between advanced AI technologies and clinical practice, ultimately contributing to the evolution and integration of intelligent healthcare systems [19, 20, 14, 21, 1].

1.3 Importance of Integrating Multiple Data Modalities

Integrating multiple data modalities in MLLMs is essential for advancing clinical applications by providing a holistic understanding of complex medical scenarios. The combination of structured clinical notes with unstructured data from electronic health records (EHRs) facilitates the synthesis of diverse data types, significantly improving the accuracy of medical predictions and enhancing triage recommendations. This approach leverages rich information contained in clinical notes—often overlooked in traditional models—alongside structured data such as vital signs and laboratory results. Research demonstrates that models incorporating both data types outperform those relying solely on structured data, highlighting the critical role of comprehensive data integration in advancing healthcare outcomes [22, 11]. The necessity for processing unstructured clinical notes underscores the transformative potential of MLLMs in healthcare delivery.

The integration of domain-specific visual datasets with vision-language datasets is pivotal for enhancing MLLM capabilities, particularly in applications like visual question answering [4]. Methods such as MedRegA, which incorporate region-centric information, emphasize the importance of multimodal integration for improved interpretative capabilities [5]. Moreover, combining LLMs with visual encoders enhances multimodal understanding, improving performance across various tasks [8].

To enhance representation learning in medical foundation models, integrating features at local, instance, modality, and global scales is essential [23]. This comprehensive approach addresses limitations of existing methods that hinder small-scale MLLMs from effectively learning from large-scale models due to capacity constraints [9]. Additionally, integrating secure data sharing mechanisms with advanced retrieval techniques is crucial for enhancing MLLM outputs in healthcare [7].

The development of benchmarks that incorporate fine-grained concept annotations further illustrates the critical role of multimodal integration. These benchmarks deepen understanding of visual and textual information, which is essential for the enhanced performance of MLLMs in clinical settings [6]. By leveraging multiple data modalities, MLLMs can deliver more accurate and comprehensive insights, ultimately leading to improved diagnostic accuracy and personalized treatment plans in healthcare environments.

1.4 Structure of the Survey

This survey is systematically organized into seven main sections to comprehensively explore MLLMs in healthcare. The introductory section establishes a foundational understanding of MLLMs, emphasizing their significance and transformative potential in healthcare applications while outlining the scope and objectives, particularly the importance of integrating multiple data modalities to enhance clinical applications.

The second section delves into background concepts, defining key terms such as multimodal data and large language models, and discussing the evolution of MLLMs within healthcare. The third section addresses the technological foundations of MLLMs, exploring techniques for integrating various data modalities, architectural innovations, training methodologies, and the role of pre-trained models and frameworks.

In the fourth section, the survey examines clinical applications of MLLMs, highlighting their contributions to diagnostic accuracy, personalized treatment plans, and patient care, particularly in resource-constrained environments. The fifth section identifies challenges and limitations associated with implementing MLLMs in healthcare, including data challenges, computational constraints, model limitations, and ethical considerations.

The sixth section explores future directions and opportunities for MLLMs, discussing ongoing research, potential advancements, and integration with emerging technologies. It emphasizes the significance of collaborative research and clinical validation in advancing MLLMs.

In conclusion, the findings encapsulate the transformative potential of MLLMs in healthcare, emphasizing their ability to integrate diverse data types—text, images, and audio—to enhance clinical decision support, patient engagement, and medical research. The survey also highlights challenges associated with MLLM implementation, including data limitations and ethical considerations, while identifying critical research gaps that need addressing. This underscores the necessity for ongoing research and collaborative efforts to ensure the responsible integration of MLLMs into medical practice, ultimately driving innovation and improving patient outcomes in the healthcare sector [1, 19, 3].The following sections are organized as shown in Figure 1.

2 Background and Core Concepts

2.1 Definitions and Key Concepts

Multimodal Large Language Models (MLLMs) represent a significant advancement in AI, designed to process and integrate diverse data types such as text, images, and videos to improve healthcare delivery. These models effectively interpret structured data, often via machine learning, and unstructured data, like clinical notes, through natural language processing [5]. The integration of imaging, molecular diagnostics, and electronic health records (EHRs) is fundamental to MLLM functionality [8].

A key feature of MLLMs is their "multimodal comprehension," which merges visual and textual data to provide a comprehensive understanding of complex medical information. This capability is crucial for pixel-level analysis in medical imaging. However, the computational demands of processing high-dimensional data present challenges, necessitating optimization for effective clinical application [6].

Order sensitivity in MLLMs is critical, as the sequence of multimodal inputs can significantly impact model performance. The multi-granularity noisy correspondence (MNC) problem highlights inaccuracies that can arise during retrieval and generation, necessitating robust evaluation methods to assess MLLM capabilities [24]. Benchmarks like SEED-Bench-2-Plus have been developed to tackle the challenge of extracting and grounding information from multiple modalities, including text, image, audio, and video [10].

Security challenges, such as clinical mismatches and malicious queries in medical QA tasks, are addressed through specialized benchmarks that introduce attack scenarios, including the 2M-attack and its optimized variant, the O2M-attack. These expose vulnerabilities in MLLMs, revealing potential risks to patient safety and the efficacy of medical AI applications [25, 26]. By defining and addressing these key concepts, MLLMs are poised to enhance healthcare delivery through improved data integration and interpretation, while effectively managing security and privacy concerns.

2.2 Evolution of MLLMs in Healthcare

The evolution of MLLMs in healthcare has been driven by the need to extend AI interpretative capabilities beyond traditional text-based models. Early efforts aimed to overcome limitations of text-centric language models in the complex healthcare domain [14]. The integration of domain-specific visual information has become essential for accurate medical diagnoses and treatment planning [4].

A pivotal advancement was the introduction of Med-2E3, which incorporated both 3D and 2D features for enhanced medical image analysis, significantly improving performance in complex image interpretation [27]. This underscored the importance of multidimensional data integration for precision in medical imaging. Concurrently, models like EE-MLLM exemplify the shift towards optimizing data and computational efficiency, addressing challenges of data scarcity and resource constraints [10].

Historically, the development of MLLMs has focused on mitigating task interference from integrating diverse data modalities. While earlier approaches primarily addressed textual differences, recent methodologies emphasize visual information, expanding MLLM scope to encompass comprehensive data interpretation [24]. The introduction of benchmarks such as MMGIC, with structured templates for multi-grained concept annotations, illustrates the evolution towards holistic data synthesis and analysis [6].

The ongoing evolution of MLLMs in healthcare reflects a commitment to enhancing their functionalities, enabling these models to proficiently process and integrate diverse data types—text, images, and audio—thereby improving clinical decision support, patient engagement, and research capabilities. This development addresses challenges such as data limitations and ethical considerations, crucial for meeting the complex demands of modern healthcare and ultimately improving diagnostic accuracy and personalized patient care [28, 1, 19, 3].

3 Technological Foundations of MLLMs

Category	Feature	Method
Multimodal Data Integration Techniques	Feature Aggregation and Integration Attention and Focus Mechanisms Scale and Detail Processing	VQA-IN[4], HRMMF[7] EE-MLLM[10], MedRegA[5] MSFL[23]
Architectural Innovations	Visual Processing Enhancements Multimodal Integration Efficiency and Optimization Attention Mechanisms	MIVL[29], MM[30] AR[31], MLLM-Tool[32], OLLM[33] MCV[34] WINGS[35]
Training Methodologies and Strategies	Adaptive Techniques	MoME[24], ADRL[36], LLM-Triage[37], MM1.5[38]
Role of Pre-trained Models and Frameworks	Medical Knowledge Enhancement	M2E3[27], MKA[39]

Table 1: This table provides a comprehensive overview of the methods and techniques employed in advancing Multimodal Large Language Models (MLLMs) within the healthcare domain. It categorizes these methodologies into multimodal data integration techniques, architectural innovations, training methodologies, and the role of pre-trained models and frameworks, highlighting specific methods and frameworks that enhance data integration, model efficiency, and interpretative capabilities.

Category	Feature	Method
Multimodal Data Integration Techniques	Feature Aggregation and Integration Attention and Focus Mechanisms Scale and Detail Processing	VQA-IN[4], HRMMF[7] EE-MLLM[10], MedRegA[5] MSFL[23]
Architectural Innovations	Visual Processing Enhancements Multimodal Integration Efficiency and Optimization Attention Mechanisms	MIVL[29], MM[30] AR[31], MLLM-Tool[32], OLLM[33] MCV[34] WINGS[35]
Training Methodologies and Strategies	Adaptive Techniques	MoME[24], ADRL[36], LLM-Triage[37], MM1.5[38]
Role of Pre-trained Models and Frameworks	Medical Knowledge Enhancement	M2E3[27], MKA[39]

Table 2: This table provides a comprehensive overview of the methods and techniques employed in advancing Multimodal Large Language Models (MLLMs) within the healthcare domain. It categorizes these methodologies into multimodal data integration techniques, architectural innovations, training methodologies, and the role of pre-trained models and frameworks, highlighting specific methods and frameworks that enhance data integration, model efficiency, and interpretative capabilities.

Table 5 provides a detailed summary of the various methods and innovations that underpin the development of Multimodal Large Language Models (MLLMs) in healthcare, emphasizing their role in improving data integration and processing capabilities. The effective integration of text, images, and audio is vital for enhancing Multimodal Large Language Models (MLLMs), particularly in healthcare. This capability allows MLLMs to derive comprehensive insights from electronic health records, medical imaging, and wearable sensors, thereby addressing the complexities inherent

in modern medical data analysis [28, 1, 3]. By employing diverse multimodal data integration techniques, MLLMs enhance their interpretative capabilities and understanding of complex clinical data. Table 2 offers a detailed summary of the various methods and innovations that underpin the development of Multimodal Large Language Models (MLLMs) in healthcare, emphasizing their role in improving data integration and processing capabilities.

Figure 2 illustrates the technological foundations of MLLMs, categorized into multimodal data integration techniques, architectural innovations, training methodologies, and the role of pre-trained models and frameworks. Each category highlights specific frameworks, methods, and strategies that enhance MLLMs' capabilities in healthcare, improving data integration, model efficiency, and interpretative abilities. This figure serves to visually encapsulate the multifaceted approaches that underpin MLLMs, thereby reinforcing the discussion on their significance in advancing healthcare analytics.

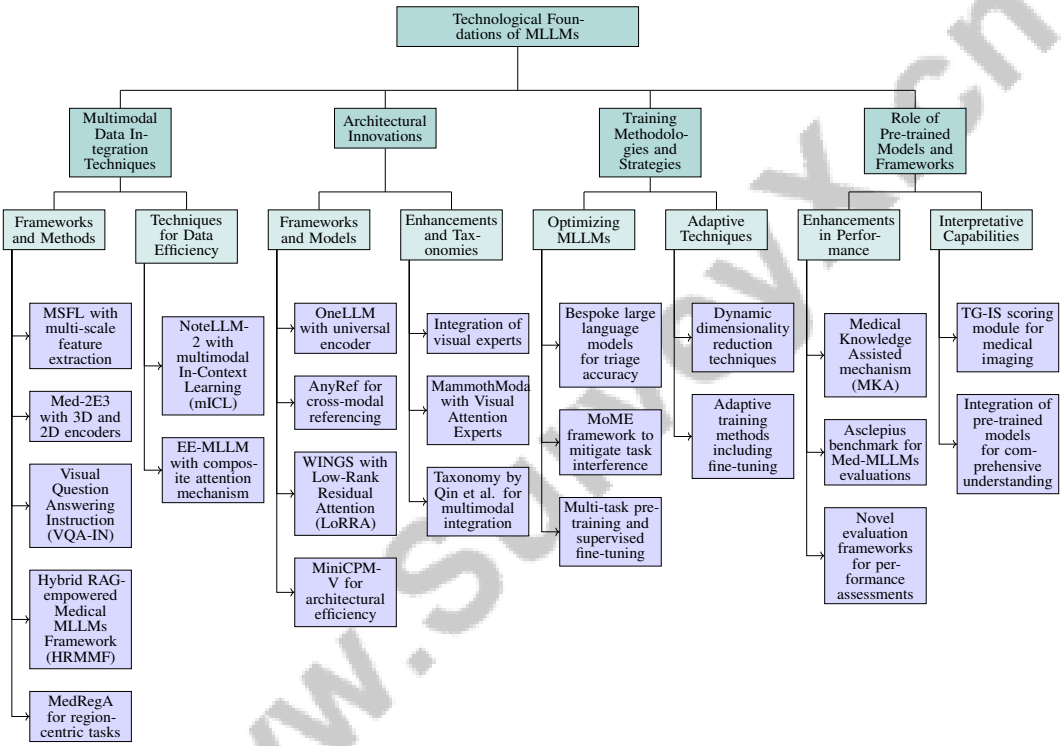


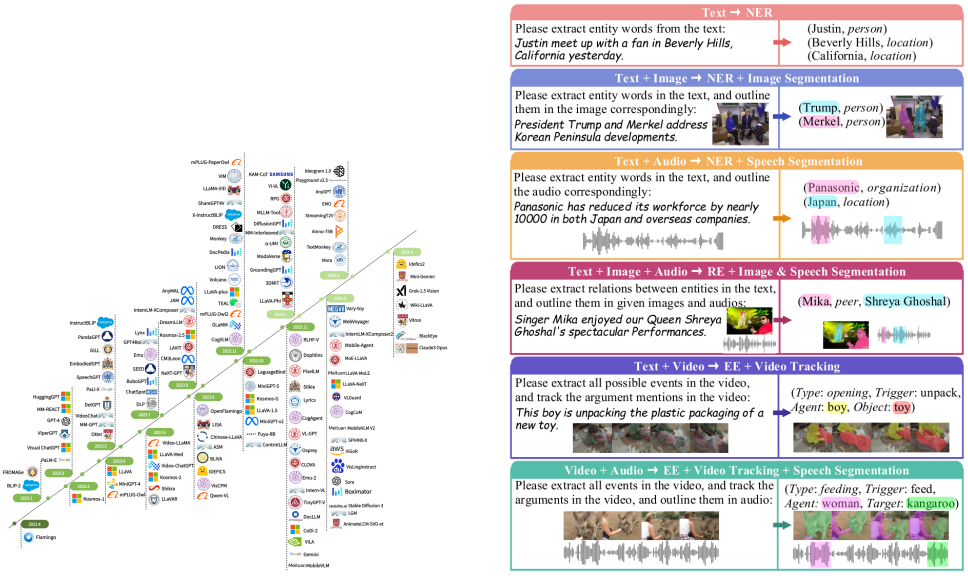
Figure 2: This figure illustrates the technological foundations of Multimodal Large Language Models (MLLMs), categorized into multimodal data integration techniques, architectural innovations, training methodologies, and the role of pre-trained models and frameworks. Each category highlights specific frameworks, methods, and strategies that enhance MLLMs' capabilities in healthcare, improving data integration, model efficiency, and interpretative abilities.

3.1 Multimodal Data Integration Techniques

The integration of various data modalities is crucial for advancing the analytical capabilities of MLLMs in healthcare. Innovative techniques have emerged to facilitate this integration, enhancing MLLMs' clinical applications. The MSFL framework, which uses multi-scale feature extraction, improves medical models' performance across clinical tasks by capturing features at different scales [23]. Med-2E3 employs both 3D and 2D encoders for feature extraction from medical images and task instructions, using a text-guided inter-slice scoring module for enhanced feature aggregation [27]. The Visual Question Answering Instruction (VQA-IN) method transforms datasets into a question-answering format, effectively unifying visual and textual information [4].

The Hybrid RAG-empowered Medical MLLMs Framework (HRMMF) combines hybrid RAG techniques with contract theory for secure healthcare data management [7]. MedRegA enhances integration through region-centric tasks, improving the model's ability to identify and report on

specific anatomical regions [5]. NoteLLM-2 employs multimodal In-Context Learning (mICL) and late fusion for embedding integration, highlighting the importance of detailed annotations [8, 6]. The EE-MLLM method uses a composite attention mechanism to enhance data efficiency, addressing the challenges of processing large multimodal data volumes [10]. These techniques collectively enable MLLMs to synthesize diverse data types, improving diagnostic accuracy and patient care.



(a) Timeline of Major Language Model Projects and Their Release Dates[40]

(b) Text-based Entity Extraction and Annotation[41]

Figure 3: Examples of Multimodal Data Integration Techniques

Figure 3 illustrates the exploration of multimodal data integration techniques. The timeline image showcases the evolution of significant language model projects, while the second image highlights the sophistication of current language models in integrating diverse data types for tasks such as named entity recognition and image segmentation [40, 41].

3.2 Architectural Innovations

Method Name	Integration Techniques	Efficiency Enhancements	Representation Learning
OLLM[33]	Dynamic Routing Mechanism	Universal Encoder Projection	Universal Encoder Projection
AR[31]	Refocusing Mechanism	Attention Scores	Unified Representation
WINGS[35]	Visual Learners Integration	Low-Rank Residual	Parallel Learners
MCV[34]	Visual Encoder Integration	Optimized Deployment Strategies	Adaptive Visual Encoding
MIVL[29]	Mixture-of-experts	Delta Tuning	Endogenous Visual Pre-training
MM[30]	Visual Merger Module	Architectural Optimizations	Vision Encoder
MLLM-Tool[32]	Multi-modal Encoders	Low-Rank Adaptation	Linear Projection Layer

Table 3: Overview of architectural innovations in multimodal large language models, highlighting the integration techniques, efficiency enhancements, and representation learning strategies employed by various methods. This table provides a comparative analysis of different frameworks, showcasing their unique contributions to improving model efficiency and cross-modal data processing capabilities.

Architectural innovations are pivotal for MLLMs' ability to process and integrate diverse data modalities effectively. Table 3 presents a detailed comparison of recent architectural innovations in multimodal large language models, emphasizing the diverse strategies employed for integration, efficiency, and representation learning. The OneLLM framework, utilizing a universal encoder and projection module, enhances model efficiency compared to traditional methods [33]. Guarrasi et al. categorize current methods based on modality types, architecture, and fusion strategies, emphasizing feature extraction and representation learning [42].

The AnyRef framework introduces a unified representation for cross-modal referencing, processed uniformly by the LLM to generate grounded outputs [31]. WINGS architecture employs Low-Rank

Residual Attention (LoRRA) to enhance visual and textual data integration [35]. MiniCPM-V emphasizes architectural efficiency through advanced designs and optimization techniques [34]. Luo et al. integrate visual experts into the LLM, facilitating effective visual learning while retaining pre-trained language knowledge [29]. The MammothModa architecture enhances multimodal processing with Visual Attention Experts and a Visual Merger Module [30].

Qin et al. introduced a taxonomy categorizing research based on data contributions to models and vice versa, highlighting the interdependence of architectural frameworks in multimodal integration [28]. ImageBind supports multi-modal encoding, showcasing architectural innovations that drive MLLM advancements [32]. These innovations collectively improve MLLMs' ability to integrate diverse data types, enhancing their applicability in healthcare.

3.3 Training Methodologies and Strategies

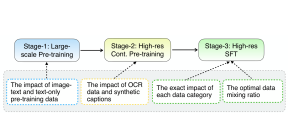
Method Name	Training Techniques	Data Handling	Efficiency Optimization
LLM-Triage[37]	Bespoke Models	Data Heterogeneity	Dynamic Dimensionality Reduction
MoME[24]	Adaptive Training Methods	Instance-level Routers	Dynamic Adaptation
ADRL[36]	Adaptive Training Methods	Data Heterogeneity Strategies	Dynamic Dimensionality Reduction
MM1.5[38]	Three-stage Training	Data Curation	Mixture-of-experts

Table 4: Overview of various training methodologies and strategies employed in Multimodal Large Language Models (MLLMs) for optimizing data integration and processing. The table highlights specific methods, including LLM-Triage, MoME, ADRL, and MM1.5, detailing their training techniques, data handling approaches, and efficiency optimization strategies.

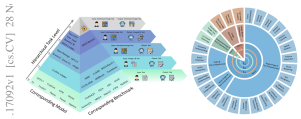
Training methodologies are essential for optimizing MLLMs' integration and processing of diverse data modalities in healthcare. Bespoke large language models enhance triage accuracy by capturing intricate language patterns from unstructured clinical notes [37]. The MoME framework mitigates task interference by combining multiple vision and language experts, enhancing performance across tasks [24]. Multi-task pretraining and supervised fine-tuning, as in MammothModa, develop robust multimodal representations [**?**]. Addressing data heterogeneity and class imbalance is crucial for managing healthcare data complexities [11].

Dynamic dimensionality reduction techniques, such as ADRL, enhance computational efficiency and accuracy by optimizing input data dimensionality [36]. This adaptation is crucial for managing the computational demands of processing large-scale multimodal data. While many approaches utilize textual detection information without training, adaptive training methods, including fine-tuning pre-trained MLLMs, significantly improve performance by enhancing comprehension of formatted textual data [28, 43]. These strategies enhance MLLMs' interpretative and analytical capabilities, improving healthcare delivery and clinical insights.

Table 4 provides a comprehensive overview of the training methodologies and strategies utilized in Multimodal Large Language Models (MLLMs) to enhance their performance in processing diverse data modalities within the healthcare domain.



(a) Large-scale Pre-training for High-Resolution Image Captioning[38]



(b) Deep Learning Models for Image-Text Interleaving and Their Applications[44]

(c) Fine-tuning Based Infusion (FTBI) for Object Detection[43]

Figure 4: Examples of Training Methodologies and Strategies

Figure 4 presents various training methodologies that enhance MLLMs' capabilities. The approaches include structured pre-training for high-resolution image captioning, deep learning models for image-text interleaving, and fine-tuning techniques for object detection [38, 44, 43].

3.4 Role of Pre-trained Models and Frameworks

Pre-trained models and frameworks are crucial for MLLMs' performance in processing diverse data modalities. These models leverage extensive datasets and architectures to capture complex patterns, addressing challenges in multimodal integration and enabling accurate complex reasoning tasks [45, 46]. Frameworks like the Medical Knowledge Assisted mechanism (MKA) embed medical knowledge graphs into neural generative models, enhancing contextual understanding [39]. Benchmarks such as Asclepius ensure the integrity of Med-MLLMs evaluations, utilizing medical materials and quizzes [47]. Novel evaluation frameworks introduce metrics for quantifying MLLMs' susceptibility to leading questions, refining performance assessments [48].

The TG-IS scoring module, mimicking radiologists' attention mechanisms, exemplifies how pre-trained frameworks enhance MLLMs' interpretative capabilities in medical imaging [27]. By integrating pre-trained models and frameworks, MLLMs achieve a comprehensive understanding of complex medical information, improving clinical outcomes and patient care. The ongoing refinement of these foundational elements is vital for advancing MLLMs' capabilities in healthcare.

Feature	MSFL	Med-2E3	VQA-IN
Integration Technique	Multi-scale Feature Extraction	3D And 2D Encoders	Question-answering Format
Architectural Innovation	Not Specified	Not Specified	Not Specified
Training Strategy	Not Specified	Not Specified	Not Specified

Table 5: This table presents a comparative analysis of three multimodal data integration techniques employed in Multimodal Large Language Models (MLLMs) within the healthcare domain. It highlights the specific integration techniques utilized by each method, namely MSFL, Med-2E3, and VQA-IN, while noting the absence of specified architectural innovations and training strategies. This comparison underscores the diverse approaches to enhancing data processing and interpretative capabilities in complex clinical environments.

4 Clinical Applications of MLLMs

Multimodal Large Language Models (MLLMs) are pivotal in transforming clinical applications by integrating diverse data sources to enhance diagnostic precision and personalize patient care. This section explores their impact on diagnostic accuracy and personalized treatment plans.

4.1 Enhancing Diagnostic Accuracy

MLLMs enhance diagnostic accuracy by synthesizing structured and unstructured data, improving clinical outcome predictions. For instance, a bespoke LLM-based approach has refined diagnostic accuracy in mental healthcare by utilizing clinical notes [37]. The Med-2E3 model excels in 3D medical image analysis, outperforming existing models by 14

Models like MedRegA enhance diagnostic precision by accurately identifying regions in medical images [5]. Multi-scale feature learning improves representation capabilities across clinical tasks [23]. The EE-MLLM model demonstrates high benchmark performance with computational efficiency, contributing to diagnostic accuracy [10]. The MoME framework mitigates task interference, enhancing generalist MLLMs' performance in diverse tasks [24]. Furthermore, MMGIC experiments show that models trained with multi-grained concept annotations outperform those using coarse-grained annotations, leading to significant benchmark improvements [6].

Challenges persist in integrating diverse modalities due to modality bias, necessitating further research for optimal performance. MLLMs are pivotal in transforming healthcare diagnostics by integrating text, images, and audio into cohesive frameworks that enhance clinical decision support and patient engagement. Understanding MLLMs' potential and limitations is essential for their responsible implementation in healthcare [1, 3].

4.2 Personalized Treatment Plans

MLLMs are instrumental in developing personalized treatment plans by integrating diverse data modalities to tailor medical interventions. By synthesizing information from electronic health records

(EHRs), medical imaging, and genomic data, MLLMs create comprehensive patient profiles that inform individualized treatment strategies [37]. This capability is crucial in precision medicine, where treatments are customized based on genetic makeup and clinical history.

MLLMs excel in processing unstructured clinical notes, extracting nuanced insights critical for formulating individualized treatment plans [5]. Their integration of visual and textual data allows for the identification of specific biomarkers essential for effective therapeutic approaches [4]. For instance, the Med-2E3 model enhances interpretative accuracy in 3D medical images, guiding personalized treatment decisions [27].

Advanced feature extraction techniques, such as those in the MSFL framework, facilitate the identification of patient-specific risk factors, optimizing therapeutic outcomes [23]. MLLMs' ability to process multimodal data enhances their effectiveness in developing personalized treatment plans, fostering a holistic understanding of patient health [10].

Secure data management techniques, as seen in the HRMMF framework, ensure sensitive patient information is handled with care while enabling personalized treatment plan generation [7]. Additionally, multi-grained concept annotations improve MLLMs' precision in tailoring treatments, providing deeper insights into complex medical data [6].

Integrating MLLMs into personalized medicine represents a significant leap in healthcare delivery, as these advanced AI systems synthesize diverse data types to create tailored interventions. This innovative approach enhances patient outcomes by providing comprehensive insights into individual health profiles, optimizing treatment strategies, and encouraging collaboration across medical disciplines while addressing data privacy and ethical considerations [1, 19, 49, 3]. By leveraging MLLMs' comprehensive data synthesis capabilities, healthcare providers can develop more effective treatment plans, ultimately enhancing patient care quality.

4.3 Improving Patient Care

MLLMs enhance patient care by integrating diverse data modalities to provide comprehensive healthcare insights. These models improve clinical decision-making and patient management by synthesizing structured data, such as EHRs, with unstructured data, including clinical notes and medical imaging. The DF-DM model exemplifies this approach, achieving high performance in healthcare applications through advanced multimodal data processing techniques, contributing to patient care improvements [50].

By leveraging MLLMs, healthcare providers can offer personalized and effective care. The integration of visual and textual data allows for a deeper understanding of patient conditions, enabling timely interventions and reducing complications. MLLMs enhance diagnostic accuracy and treatment planning by simultaneously processing and analyzing complex medical images and textual data, providing comprehensive insights into patient health and improving clinical decision support and patient engagement. Ongoing refinement of MLLM training methods, including adaptive strategies, optimizes their ability to interpret intricate visual elements, further supporting better patient outcomes [3, 43, 51, 52, 1].

Additionally, MLLMs play a crucial role in monitoring and follow-up processes, analyzing patient data over time to identify trends and predict potential health issues. This proactive approach facilitates early intervention and continuous care management, enabling personalized care tailored to evolving health statuses. By integrating diverse data sources and employing predictive modeling, this strategy enhances decision-making processes, ultimately improving patient outcomes and fostering a shift toward a preventive and patient-centered healthcare model [17, 53, 49, 54].

Integrating MLLMs into healthcare signifies a significant advancement in patient care, as these innovative tools leverage diverse data types to enhance clinical decision-making, improve diagnostic accuracy, and facilitate personalized treatment plans. By synthesizing vast amounts of information from EHRs, medical imaging, and wearable sensors, MLLMs can substantially improve clinical outcomes while addressing challenges such as data limitations and ethical considerations in their implementation [17, 1, 3]. By providing healthcare professionals with comprehensive insights, MLLMs contribute to informed decision-making and optimized patient management strategies.

4.4 Applications in Resource-Constrained Environments

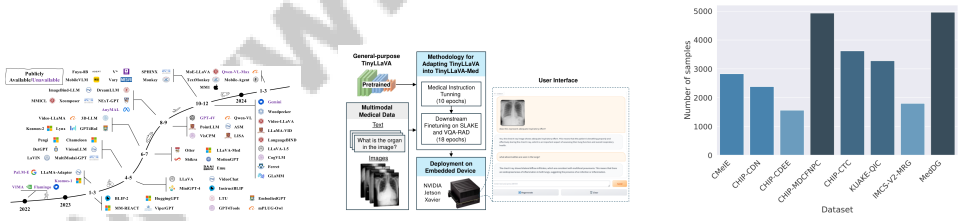
Deploying MLLMs in resource-constrained environments presents unique challenges and opportunities, necessitating strategies that optimize performance while minimizing computational demands. The TinyLLaVA-Med model exemplifies a compact MLLM designed for medical applications, optimized for efficient operation on hardware-constrained devices [55]. This model demonstrates the feasibility of advanced AI systems in settings with limited resources, broadening access to cutting-edge healthcare technologies.

Vector embeddings significantly democratize multimodal deep learning in low-resource settings, enhancing AI adaptability across diverse healthcare applications [56]. By leveraging these embeddings, MLLMs can effectively process and integrate diverse data modalities, ensuring healthcare providers in resource-limited settings benefit from AI advancements.

The LLaVA-KD framework illustrates the potential of MLLMs in resource-constrained environments by maintaining high performance while reducing model size and computational complexity [9]. This framework is suitable for settings with limited resources, enabling robust AI systems' deployment without performance compromise. Similarly, MiniCPM-V showcases strong performance across benchmarks while being deployable on end-side devices [34], highlighting the importance of balancing performance with resource efficiency.

Innovations such as adapting compressed image latents for MLLM-based vision tasks further enhance performance in resource-constrained environments [57]. Optimizing data representation allows these models to deliver high-quality outputs without extensive computational resources. The SPIDER framework effectively mitigates catastrophic forgetting in MLLM fine-tuning, enhancing specialization performance while preserving generalization, crucial for broader multimodal task applications [58].

The application of MLLMs in resource-constrained environments underscores the need for efficient models that sustain high performance while significantly reducing computational demands. This is vital as MLLMs manage and integrate various data modalities, often requiring complex task decomposition into subtasks processed by multiple pre-trained models. Researchers are optimizing MLLMs through techniques like knowledge-enhanced reranking and noise-injected training, ensuring effectiveness in dynamic contexts while minimizing reliance on extensive computational resources [59, 51]. By leveraging innovative strategies and frameworks, MLLMs can be effectively deployed in diverse healthcare settings, making advanced AI-driven solutions accessible to a wider range of healthcare providers and patients.



(a) Timeline of Publicly Available/Unavailable AI Models[60]

(b) Adapting a General-Purpose Model for Medical Image Analysis[55]

(c) Comparison of Sample Sizes Across Different Datasets[61]

Figure 5: Examples of Applications in Resource-Constrained Environments

As shown in Figure 5, exploring MLLMs' clinical applications in resource-constrained environments reveals their transformative potential in healthcare, particularly in settings with limited resources. The timeline of publicly available and unavailable AI models from 2022 to 2024 illustrates the rapid evolution and accessibility of AI technologies, highlighting key developments such as the introduction of Fuyu-8B in 2022. This progression underscores the growing availability of AI tools for medical purposes. The adaptation of general-purpose models like TinyLLaVA for specific tasks, such as medical image analysis, exemplifies innovative methodologies enhancing diagnostic capabilities. By pre-training on extensive datasets and fine-tuning for medical applications, these models are optimized for deployment on cost-effective embedded devices like the NVIDIA Jet, making advanced medical imaging solutions feasible even in limited technological infrastructure. Furthermore, comparing

sample sizes across datasets illustrates the diversity and scale of data available for training these models, ensuring robust and accurate performance across various medical scenarios. Collectively, these examples highlight MLLMs' potential to democratize healthcare by providing sophisticated AI-driven solutions accessible in resource-constrained environments [60, 55, 61].

5 Challenges and Limitations

5.1 Data Challenges

The implementation of Multimodal Large Language Models (MLLMs) in healthcare is significantly impeded by issues related to data acquisition, quality, and integration. A major challenge is the lack of comprehensive, high-quality datasets that adequately reflect the complexity and diversity of clinical scenarios, which is crucial for ensuring MLLMs' generalizability and robustness [37]. The reliance on publicly available datasets often leads to data leakage, undermining model evaluations and hindering thorough assessments of MLLM capabilities [14]. In scenarios with limited data or tasks diverging from training contexts, data acquisition challenges become pronounced, affecting MLLM implementation [27]. Moreover, the completeness and quality of data, particularly clinical notes, are critical for accurate model predictions [37], while the lack of detailed annotated datasets for domain-specific visual tasks complicates training and limits customization [4].

Integrating comprehensive medical knowledge into MLLMs remains a challenge, as models often fail to incorporate such knowledge effectively [8]. Scenarios where intermodal correlations are weakly established can also affect performance [23]. Ensuring data transparency and navigating real-world implementation complexities present further hurdles [7]. To enhance MLLM capabilities, developing comprehensive benchmarks and datasets that address data quality, diversity, and integration of various data types is essential. These advancements will improve model performance in clinical decision support and patient engagement while addressing challenges related to data privacy and annotation costs [59, 1, 48, 62]. Overcoming these data-related challenges is crucial for MLLMs to provide accurate and reliable clinical outcomes, ultimately enhancing healthcare delivery and patient care.

5.2 Computational and Resource Constraints

The deployment of MLLMs in healthcare is significantly limited by computational and resource constraints, affecting the feasibility and efficiency of these models in real-world applications. The substantial computational cost associated with running MLLMs, especially those with numerous parameters, limits their use in mobile, offline, energy-sensitive, and privacy-protective scenarios, compounded by the need for extensive server-side resources [24]. Challenges also arise from reliance on specific datasets that may not encompass all real-world retrieval scenarios, potentially affecting MLLM generalizability across diverse clinical settings [6]. Imbalances in training data distribution among modalities can further impact model performance, particularly when certain modalities are underrepresented [5].

High-dimensional data necessitates careful tuning of hyperparameters for optimal performance, and the dependency on segmentation quality, as seen in methods like SAM, may constrain MLLM performance in clinical applications [36]. Models may also struggle with extremely high-resolution inputs and long-duration videos, straining computational resources. Innovative solutions, such as the efficient adaptation of compressed latents, offer promising approaches to address these constraints without extensive computational resources. However, existing methods often fail to sufficiently integrate domain-specific knowledge into medical LLMs, limiting their ability to handle complex medical consultations [63]. Addressing computational and resource limitations is crucial for MLLMs to integrate diverse data types and deliver comprehensive insights into patient health, requiring innovative strategies that optimize resource usage while maintaining performance [1, 3].

5.3 Model Limitations and Interpretability

MLLM deployment in healthcare is challenged by inherent model limitations and interpretability issues. Hallucinations and errors in pixel-level grounding, particularly for small targets, compromise model reliability in clinical applications [64]. MLLMs often struggle with fine-grained understanding and maintaining temporal coherence in video generation, vital for accurate medical image interpretation [65]. Despite advancements, models like MR-MLLM face challenges in novel object detection

and generalization to unseen categories, limiting their applicability in diverse clinical scenarios [66]. Reliance on single data sources, as seen in AITQE, restricts findings' generalizability across diverse datasets, a significant limitation given healthcare data's multifaceted nature [52].

Initial dataset analyses can introduce biases if data is not representative, affecting MLLM output interpretability and accuracy [36]. MLLMs' susceptibility to hallucinations and poor performance in complex reasoning tasks highlights the need for robust evaluation frameworks to ensure reliable model outputs [14]. Addressing these limitations, such as data constraints, technical challenges, and ethical concerns, while enhancing interpretability, is crucial for effective MLLM integration into healthcare systems. This will facilitate improved clinical decision-making, patient engagement, and research outcomes [19, 3, 59, 1, 48]. Overcoming these challenges will enable MLLMs to deliver accurate, reliable, and interpretable insights, ultimately improving healthcare delivery and patient outcomes.

5.4 Ethical and Regulatory Considerations

Integrating MLLMs in healthcare requires careful consideration of ethical and regulatory issues to ensure responsible and secure application. Patient privacy and data security are critical concerns in AI system implementation, as erroneous interpretations, such as those within SERPENT-VLM, underscore the need for meticulous application to prevent adverse outcomes [2]. Ethical implications are profound, as MLLMs may inadvertently perpetuate biases present in training data, impacting medical decision-making processes [13]. Addressing safety issues is paramount to prevent patient harm and enhance the security of vision-language models (VLMs) in clinical environments. The benchmark introduced by Huang et al. emphasizes the necessity of ensuring the safety and efficacy of MedMLLMs, given the severe consequences of security breaches [25]. Dynamic safety mechanisms, such as those integrated into CUE-M, are vital for addressing ethical and regulatory considerations by adapting to instance-specific and category-specific risks in multimodal retrieval [67]. The ECSO strategy exemplifies an innovative approach to protection by assessing MLLM response safety and transforming unsafe images into text, facilitating safe response generation [68].

Ethical considerations related to data privacy and security are critical for LLM implementation in healthcare, as highlighted by Taylor et al. [37]. Enhancing privacy protection through methods such as avoiding gradient transmission in federated learning is essential for addressing these ethical concerns [69]. By addressing these ethical and regulatory considerations, healthcare providers can ensure that MLLMs are integrated into clinical practice in a manner that enhances patient care while upholding ethical standards. Robust regulatory frameworks are essential to guide the ethical use of multimodal data in clinical settings, ensuring AI systems respect patient rights and foster trust in healthcare innovations.

5.5 Clinical Validation and Real-World Application

The clinical validation and real-world application of MLLMs in healthcare are critical for ensuring efficacy and reliability across diverse clinical settings. Rigorous clinical validation is essential for confirming the accuracy, safety, and effectiveness of these advanced AI models in real-world medical environments. Such validation is crucial for integrating diverse data types and supporting various healthcare applications, from clinical decision-making to patient engagement, thereby overcoming the technical, ethical, and operational hurdles that hinder widespread adoption [70, 15, 16, 71, 1].

A significant obstacle to MLLMs' widespread adoption in real-world healthcare applications is the lack of efficient models capable of operating within limited computational resources available in remote and resource-constrained settings, restricting access to advanced diagnostic technologies necessary for comprehensive patient evaluation and management [55]. Deploying MLLMs in such environments requires models that are not only accurate but also computationally efficient and adaptable to varying resource conditions.

Moreover, the real-world application of MLLMs is challenged by the need for seamless integration with existing healthcare workflows and systems. This integration demands robust interoperability and the ability to handle diverse data types and sources, ensuring MLLMs function effectively alongside traditional healthcare technologies. Variability in data quality and availability across healthcare settings presents significant challenges for maintaining consistent performance in machine learning models. Developing adaptive models capable of learning from new data inputs over time is

essential, particularly as healthcare increasingly relies on a wide array of complex and heterogeneous data sources, such as electronic health records and medical imaging. Innovative approaches like patchwork learning can facilitate integration of disparate datasets while preserving data privacy, ultimately enhancing generalizability and effectiveness in addressing various healthcare challenges [71, 72, 11, 73].

Ongoing research and collaborative efforts are crucial for developing clinically validated MLLMs that are practical for real-world application. This includes creating comprehensive benchmarks and datasets that reflect the complexity of real-world clinical scenarios and establishing standardized evaluation frameworks to assess model performance across diverse healthcare environments. By addressing challenges related to data limitations, technical hurdles, and ethical considerations, MLLMs can be seamlessly integrated into clinical practice, enhancing healthcare delivery through improved clinical decision support, patient engagement, and comprehensive analysis of diverse data types, ultimately leading to better patient outcomes and more effective healthcare systems [1, 19, 3].

6 Future Directions and Opportunities

6.1 Development of Comprehensive Benchmarks and Datasets

Benchmark	Size	Domain	Task Format	Metric
MIA[74]	400	Image Instruction Following	Open-ended Responses	Instruction Adherence, Overall Score
NEW-BM[75]	100,000	Image Classification	Classification	Accuracy, F1-score
SEED-Bench-2[44]	24,371	Multimodal Understanding	Multiple-Choice Questions	Accuracy, CLIP Similarity Score
DrBode[76]	109,500	Medical	Question Answering	Accuracy, Completeness
EMT[77]	200,000	Image Classification	Image Classification	Accuracy
MMRel[78]	15,000	Vision-Language	Relation Understanding	Accuracy, F1-Score
MMR[48]	48,000	Visual Question Answering	Multiple-choice Questions	Misleading Rate, Robustness Accuracy
MediConfusion[79]	80,000	Radiology	Visual Question Answering	Set accuracy, Individual accuracy

Table 6: Table ef presents a detailed comparison of various benchmarks used to evaluate Multimodal Large Language Models (MLLMs) across diverse domains and task formats. Each benchmark is characterized by its size, domain of application, task format, and the metrics used for performance evaluation. This comprehensive overview aids in understanding the scope and applicability of these benchmarks in assessing MLLM capabilities.

The progress of MLLMs in healthcare relies heavily on the creation of benchmarks and datasets that accurately capture the complexity of clinical scenarios. Table 6 provides an in-depth analysis of key benchmarks essential for assessing the performance of Multimodal Large Language Models (MLLMs) in healthcare and other domains. These benchmarks are crucial for evaluating MLLM performance across multimodal tasks, ensuring their applicability in diverse medical environments. Future research should focus on developing extensive multimodal datasets that encompass electronic health records, medical imaging, and genomic data to assess the real-world applicability and robustness of MLLMs for clinical decision support and patient engagement [28, 40, 3, 51, 1].

Enhancing dataset inclusivity is essential for improving MLLM effectiveness in varied linguistic settings and their ability to manage complex interactions [14]. Future studies should refine multi-scale integration techniques to boost applications in diverse medical contexts [23]. Expanding frameworks like MoME to include additional modalities and larger datasets is vital for a thorough evaluation of MLLMs [24].

Improving instruction tuning datasets and evaluation benchmarks to assess long-context reasoning is crucial for advancing MLLM reasoning abilities [14]. Research should explore enhancing aligner mechanisms and additional modalities to further boost model capabilities [10]. By focusing on these areas, research can significantly enhance MLLM capabilities, equipping them to meet the diverse demands of modern healthcare delivery.

6.2 Advancements in Model Architecture and Training Techniques

Enhancing MLLM capabilities in healthcare requires advancements in model architecture and training techniques. Future research should broaden evaluation benchmarks to include more tasks and improve evaluation metrics' robustness. The MME benchmark, which assesses perceptual and cognitive capabilities across 14 subtasks, highlights significant areas for improvement in existing MLLMs. Specialized benchmarks like SEED-Bench-2-Plus emphasize the need for targeted evaluations in specific contexts, such as text-rich visual comprehension, guiding MLLM optimization for real-world applications [59, 21, 80, 81].

Integrating graph neural networks offers a promising avenue for enhancing knowledge extraction and prediction capabilities in MLLMs, particularly in medical applications [39]. This approach can significantly improve MLLMs' ability to process and interpret complex medical data, leading to more accurate healthcare solutions.

Research should also refine dimensionality reduction techniques, such as ADRL, and explore their applicability in complex datasets [36]. This refinement is essential for optimizing computational efficiency while maintaining high performance across diverse healthcare scenarios.

Expanding dataset size and incorporating additional multimodal tasks are critical for enhancing evaluation frameworks' robustness, as emphasized in the TPEval framework [82]. Improving dataset quality and exploring diverse domain-specific tasks will further enhance MLLMs' adaptability and performance in various clinical environments [4].

These efforts aim to refine the architectural and training paradigms of MLLMs, ensuring their efficacy and reliability in healthcare applications. By concentrating on critical areas such as dataset development, modality alignment methods, and ethical guidelines, future research can enhance MLLM functionality, enabling them to integrate diverse data types—text, images, and audio—effectively addressing the intricate requirements of modern healthcare delivery. This will allow MLLMs to play a pivotal role in clinical decision support, medical imaging, and patient engagement, ultimately transforming healthcare landscapes [59, 1, 19, 3].

6.3 Integration with Emerging Technologies

Integrating MLLMs with emerging technologies offers substantial potential for enhancing their capabilities and addressing challenges in healthcare applications. Future research should explore advanced attention mechanisms, as demonstrated by the Med-2E3 model, which improves MLLM interpretative accuracy in complex scenarios [27]. Enhancing model performance in multi-label classification tasks and expanding datasets to include diverse medical conditions are crucial for integrating MLLMs with emerging technologies, broadening their applicability and effectiveness [5].

The potential integration of additional modalities, such as audio and video, offers significant opportunities to enhance MLLM capabilities in various recommendation scenarios, expanding utility beyond traditional text and image data [8]. This integration is particularly relevant for developing comprehensive healthcare solutions that require synthesizing diverse data types to provide accurate and personalized recommendations.

Applying frameworks like Transform-Neck to video or audio coding can facilitate seamless MLLM integration with emerging technologies, enhancing adaptability across diverse data modalities. This approach expands MLLM capabilities to manage and integrate diverse data types effectively and improves performance by employing multiple pre-trained models for individual subtasks, leading to more accurate outcomes with complex multimodal datasets [59, 51].

Integrating MLLMs with emerging technologies presents a promising avenue for advancing their capabilities, ensuring they are well-equipped to handle the complex demands of modern healthcare delivery. By enhancing the integration of multiple pre-trained models for specific subtasks and incorporating adaptive training methods for better comprehension of detection information, future research can significantly boost MLLM adaptability and performance. These advancements will ultimately increase effectiveness and applicability in clinical practice, enabling MLLMs to better interpret complex multimodal data and deliver accurate results in real-world scenarios [59, 43].

6.4 Collaborative Research and Clinical Validation

The advancement of MLLMs in healthcare depends on collaborative research and clinical validation, essential for ensuring model efficacy, safety, and reliability in real-world settings. Interdisciplinary collaboration among medical professionals, data scientists, and ethicists is pivotal, fostering responsible MLLM integration into healthcare systems. Such efforts facilitate MLLM refinement to better meet clinical decision-making demands and explore additional evaluation tasks and scenarios [3].

Future research should prioritize enhancing MLLM adaptability to a broader range of clinical contexts, as evidenced by the VS-Assistant, which could be improved to handle more complex queries and tasks across various surgical scenarios [83]. This adaptability is crucial for effectively addressing the diverse challenges encountered in healthcare environments.

Collaboration among researchers, healthcare institutions, and policymakers is crucial for advancing AI-driven precision medicine, where integrative strategies are essential for leveraging MLLM potential [49]. Developing sophisticated safety reasoning frameworks and enhancing visual understanding capabilities in MLLMs are critical areas for future research, contributing to the reliability and quality of model outputs, particularly for improving pixel-level grounding capabilities [84].

Expanding datasets to include diverse medical knowledge and exploring additional evaluation metrics are vital for better assessing model performance and ensuring robustness in healthcare applications [85]. Future research could also investigate incorporating additional types of annotations, scaling datasets, and evaluating MLLM performance across a broader range of vision-language tasks [6].

Collaborative research is essential for addressing challenges posed by medical hallucinations and improving the assessment and reliability of large vision-language models (LVLMs) in healthcare settings [86]. By fostering collaborative efforts and prioritizing clinical validation, MLLMs can be effectively integrated into healthcare systems, ultimately improving patient outcomes and advancing the field of medical AI.

7 Conclusion

The exploration of Multimodal Large Language Models (MLLMs) within the healthcare sector reveals their substantial potential to revolutionize clinical practices through the integration of diverse data types. By adeptly processing structured and unstructured information, such as electronic health records, medical imaging, and genomic data, MLLMs enhance diagnostic precision, facilitate personalized treatment strategies, and improve patient care outcomes. These models are instrumental in deriving sophisticated insights from complex datasets, thereby supporting informed clinical decisions and optimizing healthcare delivery.

Despite these advancements, the implementation of MLLMs in medical settings is not without its challenges. Obstacles related to data accessibility, computational demands, model constraints, and ethical considerations must be addressed to ensure the models' effective and responsible integration into clinical environments. Progress in developing robust benchmarks, comprehensive datasets, and innovative model architectures is crucial for overcoming these barriers and augmenting the capabilities of MLLMs.

The synergy of MLLMs with cutting-edge technologies and the promotion of collaborative research initiatives are vital for advancing medical artificial intelligence. Partnerships among healthcare practitioners, data scientists, and policymakers are essential to devise solutions that mitigate current limitations and broaden the scope of MLLMs across various healthcare contexts. Emphasizing clinical validation and practical application will facilitate the seamless incorporation of MLLMs into healthcare systems, thereby transforming healthcare delivery and enhancing patient care.

This survey highlights the imperative for sustained research and interdisciplinary collaboration to fully harness the transformative potential of MLLMs in healthcare. As the field evolves, the continuous refinement and deployment of these models will be pivotal in achieving more precise, efficient, and ethically sound healthcare solutions.

References

[1] Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Junyu Liu, Benji Peng, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. From text to multimodality: Exploring the evolution and impact of large language models in medical practice, 2024.

[2] Manav Nitin Kapadnis, Sohan Patnaik, Abhilash Nandy, Sourjyadip Ray, Pawan Goyal, and Debdoot Sheet. Serpent-vlm : Self-refining radiology report generation using vision language models, 2024.

[3] Qian Niu, Keyu Chen, Ming Li, Pohsun Feng, Ziqian Bi, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Junyu Liu, et al. From text to multimodality: Exploring the evolution and impact of large language models in medical practice. *arXiv preprint arXiv:2410.01812*, 2024.

[4] Jusung Lee, Sungguk Cha, Younghyun Lee, and Cheoljong Yang. Visual question answering instruction: Unlocking multimodal large language model to domain-specific visual multitasks, 2024.

[5] Lehan Wang, Haonan Wang, Honglong Yang, Jiaji Mao, Zehong Yang, Jun Shen, and Xiaomeng Li. Interpretable bilingual multimodal large language model for diverse biomedical tasks, 2024.

[6] Xiao Xu, Tianhao Niu, Yuxi Xie, Libo Qin, Wanxiang Che, and Min-Yen Kan. Exploring multi-grained concept annotations for multimodal large language models, 2024.

[7] Cheng Su, Jinbo Wen, Jiawen Kang, Yonghua Wang, Yuanjia Su, Hudan Pan, Zishao Zhong, and M. Shamim Hossain. Hybrid rag-empowered multi-modal llm for secure data management in internet of medical things: A diffusion-based contract approach, 2024.

[8] Chao Zhang, Haoxin Zhang, Shiwei Wu, Di Wu, Tong Xu, Yan Gao, Yao Hu, and Enhong Chen. Notellm-2: Multimodal large representation models for recommendation. *arXiv preprint arXiv:2405.16789*, 2024.

[9] Yuxuan Cai, Jiangning Zhang, Haoyang He, Xinwei He, Ao Tong, Zhenye Gan, Chengjie Wang, and Xiang Bai. Llava-kd: A framework of distilling multimodal large language models, 2024.

[10] Feipeng Ma, Yizhou Zhou, Hebei Li, Zilong He, Siying Wu, Fengyun Rao, Yueyi Zhang, and Xiaoyan Sun. Ee-mllm: A data-efficient and compute-efficient multimodal large language model, 2024.

[11] Keith Feldman, Louis Faust, Xian Wu, Chao Huang, and Nitesh V. Chawla. Beyond volume: The impact of complex healthcare data on the machine learning pipeline, 2018.

[12] Chaochao Lu, Chen Qian, Guodong Zheng, Hongxing Fan, Hongzhi Gao, Jie Zhang, Jing Shao, Jingyi Deng, Jinlan Fu, Kexin Huang, Kunchang Li, Lijun Li, Limin Wang, Lu Sheng, Meiqi Chen, Ming Zhang, Qibing Ren, Sirui Chen, Tao Gui, Wanli Ouyang, Yali Wang, Yan Teng, Yaru Wang, Yi Wang, Yinan He, Yingchun Wang, Yixu Wang, Yongting Zhang, Yu Qiao, Yujiong Shen, Yurong Mou, Yuxi Chen, Zaibin Zhang, Zhelun Shi, Zhenfei Yin, and Zhipin Wang. From gpt-4 to gemini and beyond: Assessing the landscape of mllms on generalizability, trustworthiness and causality through four modalities, 2024.

[13] Shaolin Zhu, Supryadi, Shaoyang Xu, Haoran Sun, Leiyu Pan, Menglong Cui, Jiangcun Du, Renren Jin, António Branco, and Deyi Xiong. Multilingual large language models: A systematic survey, 2024.

[14] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning, 2024.

[15] Yuting He, Fuxiang Huang, Xinrui Jiang, Yuxiang Nie, Minghao Wang, Jiguang Wang, and Hao Chen. Foundation model for advancing healthcare: Challenges, opportunities, and future directions, 2024.

[16] Wasif Khan, Seowung Leem, Kyle B. See, Joshua K. Wong, Shaoting Zhang, and Ruogu Fang. A comprehensive survey of foundation models in medicine, 2025.

[17] Shuroug A Alowais, Sahar S Alghamdi, Nada Alsuhebany, Tariq Alqahtani, Abdulrahman I Alshaya, Sumaya N Almohareb, Atheer Aldairem, Mohammed Alrashed, Khalid Bin Saleh, Hisham A Badreldin, et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC medical education*, 23(1):689, 2023.

[18] Onyekachukwu R. Okonji, Kamol Yunusov, and Bonnie Gordon. Applications of generative ai in healthcare: algorithmic, ethical, legal and societal considerations, 2024.

[19] Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine, 2024.

[20] Jiaxing Huang and Jingyi Zhang. A survey on evaluation of multimodal large language models, 2024.

[21] Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, Caifeng Shan, and Ran He. Mme-survey: A comprehensive survey on evaluation of multimodal llms, 2024.

[22] Bo Yang and Lijun Wu. How to leverage multimodal ehr data for better medical predictions? *arXiv preprint arXiv:2110.15763*, 2021.

[23] Weijian Huang, Cheng Li, Hong-Yu Zhou, Jiarun Liu, Hao Yang, Yong Liang, Guangming Shi, Hairong Zheng, and Shanshan Wang. Enhancing representation in medical vision-language foundation models via multi-scale information extraction techniques, 2024.

[24] Leyang Shen, Gongwei Chen, Rui Shao, Weili Guan, and Liqiang Nie. Mome: Mixture of multimodal experts for generalist multimodal large language models, 2024.

[25] Xijie Huang, Xinyuan Wang, Hantao Zhang, Yinghao Zhu, Jiawen Xi, Jingkun An, Hao Wang, Hao Liang, and Chengwei Pan. Medical mllm is vulnerable: Cross-modality jailbreak and mismatched attacks on medical multimodal large language models, 2024.

[26] Jan Clusmann, Dyke Ferber, Isabella C. Wiest, Carolin V. Schneider, Titus J. Brinker, Sebastian Foersch, Daniel Truhn, and Jakob N. Kather. Prompt injection attacks on large language models in oncology, 2024.

[27] Yiming Shi, Xun Zhu, Ying Hu, Chenyi Guo, Miao Li, and Ji Wu. Med-2e3: A 2d-enhanced 3d medical multimodal large language model, 2024.

[28] Zhen Qin, Daoyuan Chen, Wenhao Zhang, Liuyi Yao, Yilun Huang, Bolin Ding, Yaliang Li, and Shuiguang Deng. The synergy between data and multi-modal large language models: A survey from co-development perspective, 2024.

[29] Gen Luo, Xue Yang, Wenhan Dou, Zhaokai Wang, Jiawen Liu, Jifeng Dai, Yu Qiao, and Xizhou Zhu. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training, 2024.

[30] Qi She, Junwen Pan, Xin Wan, Rui Zhang, Dawei Lu, and Kai Huang. Mammothmoda: Multi-modal large language model, 2024.

[31] Junwen He, Yifan Wang, Lijun Wang, Huchuan Lu, Jun-Yan He, Jin-Peng Lan, Bin Luo, and Xuansong Xie. Multi-modal instruction tuned llms with fine-grained visual perception, 2024.

[32] Chenyu Wang, Weixin Luo, Qianyu Chen, Haonan Mai, Jindi Guo, Sixun Dong, Xiaohua, Xuan, Zhengxin Li, Lin Ma, and Shenghua Gao. Mllm-tool: A multimodal large language model for tool agent learning, 2024.

[33] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language, 2025.

[34] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone, 2024.

[35] Yi-Kai Zhang, Shiyin Lu, Yang Li, Yanqing Ma, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. Wings: Learning multimodal llms without text-only forgetting, 2024.

[36] Zhuokun Chen, Jinwu Hu, Zeshuai Deng, Yufeng Wang, Bohan Zhuang, and Mingkui Tan. Enhancing perception capabilities of multimodal llms with training-free fusion, 2024.

[37] Niall Taylor, Andrey Kormilitzin, Isabelle Lorge, Alejo Nevado-Holgado, and Dan W Joyce. Bespoke large language models for digital triage assistance in mental health care, 2024.

[38] Haotian Zhang, Mingfei Gao, Zhe Gan, Philipp Dufter, Nina Wenzel, Forrest Huang, Dhruti Shah, Xianzhi Du, Bowen Zhang, Yanghao Li, Sam Dodge, Keen You, Zhen Yang, Aleksei Timofeev, Mingze Xu, Hong-You Chen, Jean-Philippe Fauconnier, Zhengfeng Lai, Haoxuan You, Zirui Wang, Afshin Dehghan, Peter Grasch, and Yinfei Yang. Mm1.5: Methods, analysis insights from multimodal llm fine-tuning, 2024.

[39] Ke Liang, Sifan Wu, and Jiayi Gu. Mka: A scalable medical knowledge assisted mechanism for generative models on medical conversation tasks, 2023.

[40] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. A comprehensive review of multimodal large language models: Performance and challenges across different tasks, 2024.

[41] Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction, 2024.

[42] Valerio Guarrasi, Fatih Aksu, Camillo Maria Caruso, Francesco Di Feola, Aurora Rofena, Filippo Ruffini, and Paolo Soda. A systematic review of intermediate fusion in multimodal deep learning for biomedical applications, 2024.

[43] Qirui Jiao, Daoyuan Chen, Yilun Huang, Yaliang Li, and Ying Shen. From training-free to adaptive: Empirical insights into mllms' understanding of detection information, 2024.

[44] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models, 2023.

[45] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025.

[46] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[47] Wenxuan Wang, Yihang Su, Jingyuan Huan, Jie Liu, Wenting Chen, Yudi Zhang, Cheng-Yi Li, Kao-Jung Chang, Xiaohan Xin, Linlin Shen, and Michael R. Lyu. Asclepius: A spectrum evaluation benchmark for medical multi-modal large language models, 2024.

[48] Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, and Bo Zhao. Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly, 2024.

[49] Ghizal fatima, Risala H. Allami, and Maitham G. Yousif. Integrative ai-driven strategies for advancing precision medicine in infectious diseases and beyond: A novel multidisciplinary approach, 2023.

[50] David Restrepo, Chenwei Wu, Constanza Vásquez-Venegas, Luis Filipe Nakayama, Leo Anthony Celi, and Diego M López. Df-dm: A foundational process model for multimodal data fusion in the artificial intelligence era, 2024.

[51] Zhanpeng Chen, Chengjin Xu, Yiyan Qi, and Jian Guo. Mllm is a strong reranker: Advancing multimodal retrieval-augmented generation via knowledge-enhanced reranking and noise-injected training. *arXiv preprint arXiv:2407.21439*, 2024.

[52] Han Huang, Yuqi Huo, Zijia Zhao, Haoyu Lu, Shu Wu, Bingning Wang, Qiang Liu, Weipeng Chen, and Liang Wang. Beyond filtering: Adaptive image-text quality enhancement for mllm pretraining, 2024.

[53] Fei Jiang, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma, Yilong Wang, Qiang Dong, Haipeng Shen, and Yongjun Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017.

[54] Adam Bohr and Kaveh Memarzadeh. The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in healthcare*, pages 25–60. Elsevier, 2020.

[55] Aya El Mir, Lukelo Thadei Luoga, Boyuan Chen, Muhammad Abdullah Hanif, and Muhammad Shafique. Democratizing mllms in healthcare: Tinyllava-med for efficient healthcare diagnostics in resource-constrained settings, 2024.

[56] David Restrepo, Chenwei Wu, Sebastián Andrés Cajas, Luis Filipe Nakayama, Leo Anthony Celi, and Diego M López. Multimodal deep learning for low-resource settings: A vector embedding alignment approach for healthcare applications, 2024.

[57] Chia-Hao Kao, Cheng Chien, Yu-Jen Tseng, Yi-Hsin Chen, Alessandro Gnutti, Shao-Yuan Lo, Wen-Hsiao Peng, and Riccardo Leonardi. Bridging compressed image latents and multimodal large language models, 2025.

[58] Wenke Huang, Jian Liang, Zekun Shi, Didi Zhu, Guancheng Wan, He Li, Bo Du, Dacheng Tao, and Mang Ye. Learn from downstream and be yourself in multimodal large language model fine-tuning, 2024.

[59] Yongqiang Zhao, Zhenyu Li, Feng Zhang, Xinhai Xu, and Donghong Liu. Enhancing subtask performance of multi-modal large language model, 2023.

[60] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models, 2024.

[61] Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications, 2024.

[62] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale, 2024.

[63] Chihcheng Hsieh, Catarina Moreira, Isabel Blanco Nobre, Sandra Costa Sousa, Chun Ouyang, Margot Brereton, Joaquim Jorge, and Jacinto C. Nascimento. Dall-m: Context-aware clinical data augmentation with llms, 2024.

[64] Xiaoshuang Huang, Lingdong Shen, Jia Liu, Fangxin Shang, Hongxiang Li, Haifeng Huang, and Yehui Yang. Towards a multimodal large language model with pixel-level insight for biomedicine, 2025.

[65] Hong Chen, Xin Wang, Yuwei Zhou, Bin Huang, Yipeng Zhang, Wei Feng, Houlun Chen, Zeyang Zhang, Siao Tang, and Wenwu Zhu. Multi-modal generative ai: Multi-modal llm, diffusion and beyond, 2024.

[66] Guanqun Wang, Xinyu Wei, Jiaming Liu, Ray Zhang, Yichi Zhang, Kevin Zhang, Maurice Chong, and Shanghang Zhang. Mr-mllm: Mutual reinforcement of multimodal comprehension and vision perception, 2024.

[67] Dongyoung Go, Taesun Whang, Chanhee Lee, Hwa-Yeon Kim, Sunghoon Park, Seunghwan Ji, Jinho Kim, Dongchan Kim, and Young-Bum Kim. Cue-m: Contextual understanding and enhanced search with multimodal large language model, 2024.

[68] Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T. Kwok, and Yu Zhang. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation, 2024.

[69] Jianyi Zhang, Hao Frank Yang, Ang Li, Xin Guo, Pu Wang, Haiming Wang, Yiran Chen, and Hai Li. Mllm-fl: Multimodal large language model assisted federated learning on heterogeneous and long-tailed data. *arXiv preprint arXiv:2409.06067*, 2024.

[70] Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

[71] Subhash Nerella, Sabyasachi Bandyopadhyay, Jiaqing Zhang, Miguel Contreras, Scott Siegel, Aysegul Bumin, Brandon Silva, Jessica Sena, Benjamin Shickel, Azra Bihorac, Kia Khezeli, and Parisa Rashidi. Transformers in healthcare: A survey, 2023.

[72] Bokai Cao. Broad learning for healthcare, 2018.

[73] Suraj Rajendran, Weishen Pan, Mert R. Sabuncu, Yong Chen, Jiayu Zhou, and Fei Wang. Patchwork learning: A paradigm towards integrative analysis across diverse biomedical data sources, 2023.

[74] Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. Mia-bench: Towards better instruction following evaluation of multimodal llms, 2024.

[75] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, Jing Shao, and Wanli Ouyang. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark, 2023.

[76] Pedro Henrique Paiola, Gabriel Lino Garcia, João Renato Ribeiro Manesco, Mateus Roder, Douglas Rodrigues, and João Paulo Papa. Adapting llms for the medical domain in portuguese: A study on fine-tuning and model evaluation, 2024.

[77] Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models, 2023.

[78] Jiahao Nie, Gongjie Zhang, Wenbin An, Yap-Peng Tan, Alex C. Kot, and Shijian Lu. Mmrel: A relation understanding benchmark in the mllm era, 2024.

[79] Mohammad Shahab Sepehri, Zalan Fabian, Maryam Soltanolkotabi, and Mahdi Soltanolkotabi. Mediconfusion: Can you trust your ai radiologist? probing the reliability of multimodal medical foundation models, 2024.

[80] Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension, 2024.

[81] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024.

[82] Yuxuan Xie, Tianhua Li, Wenqi Shao, and Kaipeng Zhang. Tp-eval: Tap multimodal llms' potential in evaluation by customizing prompts, 2024.

[83] Zhen Chen, Xingjian Luo, Jinlin Wu, Danny T. M. Chan, Zhen Lei, Jinqiao Wang, Sebastien Ourselin, and Hongbin Liu. Vs-assistant: Versatile surgery assistant on the demand of surgeons, 2024.

[84] Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Anderson Compalas, Dawn Song, and Xin Eric Wang. Multimodal situational safety, 2024.

[85] Qichen Ye, Junling Liu, Dading Chong, Peilin Zhou, Yining Hua, Fenglin Liu, Meng Cao, Ziming Wang, Xuxin Cheng, Zhu Lei, and Zhenhua Guo. Qilin-med: Multi-stage knowledge injection advanced medical large language model, 2024.

[86] Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. Detecting and evaluating medical hallucinations in large vision language models, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.